ESTIMATING PER-PIXEL CLASSIFICATION CONFIDENCE OF
REMOTE SENSING IMAGERS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By
Shiguo Jiang, M.S.
Graduate Program in Geography

The Ohio State University
2012

Dissertation Committee:
Professor Desheng Liu, Advisor
Professor Karl Ola Ahlqvist
Professor Darla Karin Munroe
Copyrighted by

Shiguo Jiang

2012
ABSTRACT

Spatial data quality is an important topic in geographic information sciences and remote sensing. It has drawn attention from academic community, government agencies, and industry. Although great progress has been made on the spatial quality of interval and ratio data, the spatial uncertainty of nominal and ordinal data remains problematic. Land use land cover is one of the most important nominal data, which has broad impacts on our environment. The significance of Land use land cover change (LULCC) as an environmental factor calls for studies on the spatial data quality in LULCC. Remote sensing image classification is the most common source for LULCC. Therefore, the accuracy of remote sensing image classification is especially important.

This dissertation aims to address the emerging challenge to reporting classification confidence at pixel level. The main body of this chapter is composed of four chapters: Chapter 2-5.

A comprehensive literature review on previous studies is presented in Chapter 2. Previous methods on per-pixel classification confidence are divided into three categories: classification score based method, interpolation based method, and regression based method. After the review of previous methods in three categories, the advantages and limitations of previous studies are summarized. Specifically, three imperative issues are discussed. First, previous studies have not been rigorously validated with empirical evidence. Second, the confusion of concepts in different studies is problematic. Third, there is no comprehensive comparison of the main existing methods.
Comprehensive evaluations of classification score based method and interpolation based method are presented Chapters 3 and 4 respectively. It is found that classification score based method is promising. Specifically, margin of victory derived from classification scores are most useful in predicting classification confidence. The interpolation based method performs poorly due to the interpolation effect. The assumption of spatially continuous data for interpolation method is not always satisfied in the practice.

In Chapter 5, a new method is proposed to estimate per-pixel classification confidence using margin of victory. A simple regression model is constructed to convert margin of victory into per-pixel classification confidence. The results are validated with complete coverage reference data. Several interesting findings are summarized here. First, stable estimation of PPCC can be obtained using margin of victory. Second, sample size is the most important factor influencing PPCC. The estimate of PPCC is robust when sample size is above certain level (in my case, 1860 pixels). When sample size is too small, no reliable PPCC estimates are available. The impact of sampling schemes on PPCC estimates is negligible. The estimated PPCCs are comparable across classifiers and thus indicate the performance of different classifiers.
To

Ping and Lechuan
ACKNOWLEDGMENTS

Foremost, I am deeply grateful to my advisor Professor Desheng Liu for his enormous support, continuous encouragement, and endless patience. His generous guidance and contribution have had and will have significant influence through my life. During my study at OSU, an increasing number of talented students joined Professor Liu's group, creating a rich environment that benefits everyone. In a teacher-student team, the student always benefits the most: the student gains knowledge from the teacher while the teacher burns himself to light the students. As I am also going to be a teacher, I will pass on this spirit to my students. And with generations of efforts, I hope that we can contribute to the progress of the society.

I would like to thank my dissertation committee members, Professor Karl Ola Ahlqvist, Professor Darla Karin Munroe, and Professor Peter W. Culicover (Graduate Faculty Representative), and my candidacy committee members, Professor Ningchuan Xiao and Professor Mei-Po Kwan. Their insights and encouragement helped me to finish my study and oral exam. I enjoyed the communications with them.

I am thankful to other (or past) members in Professor Liu's research group: Shanshan Cai, Chandana Gangodagamage, Weishu Gong, Jun Li, Haixia Liu, Jung-Kuan Liu, Qian Mou, Rongrong Wan, Hongshuo Wang, Adam Wehman, Zhouxin Xi, Fan Xia, Fang Zhang, Xuesong Zhang, Yibo Zhang, Xiaolin Zhu. They have helped me in various aspects. Their encouragement and discussion are precious gifts to me. Especially, I enjoyed the lunch time with Xiaolin in Room 0160 Derby Hall, and then in the remodeled Room 1066 DB. His encouragement and inspiring ideas played a significant role in my
study. From September 2012, Hongshuo joined the lunch time in Room 1066, bringing new flavors to our conversation. I will miss the lunch hours.

Besides members in Professor Liu's group, Emily Scarborough, Lili Wang, Brian Williams, and two faculty members, Alvaro Montenegro and Joel Wainwright, have joined my rehearsals for the oral exam. Their comments and suggestions are greatly appreciated.

During the past five and half years, I stayed four years in Room 1070 DB, and the last one and half years in Room 1083. I had a great time with the many office mates. Especially, I enjoyed the time with Meng Guo in Room 1083. We came to OSU in the same year, and graduated at the same time. I am also thankful to Wei Chen and Yibo Zhang, who also came to Columbus at the same year with me. The generous help from them make my life easier.

I extend my thanks to my friends in other state and in China. They have touched my life in various ways. Especially, I want to thank Yuanjing Qi and Xiang Luo, Professor Yu Liu and Professor Fahui Wang for their encouragement and help.

I also want to thank Mr. and Mrs. Charlie and Carolyn Pickard for their parent-like love, and Carolyn for her precious editorial help.

I want to recognize the generous financial support from The Climate, Water, and Carbon Program at OSU.

Finally, I devote my deepest gratitude to my wife Ping Huang, my son Lechuan, and parents and parents in law for their great support and love. Ping has taken over all the household work. Lechuan lights my life. I also grew with Lechuan. My parents and parents in law always care about us with their deep love.
VITA

2001 .............................................................................. B.S. Geography, Peking University
2001 .............................................................................. B.A. Economics, Peking University
2004 .............................................................................. M.S. Geography, Peking University
2004-2007 ........................................................ Urban Planner, Beijing Tsinghua Urban Planning and Design Institute
2007-2012 ........................................................ Graduate Research Associate, Department of Geography, The Ohio State University
2011, 2012 (Winter Quarter) ........................ Graduate Teaching Associate, Department of Geography, The Ohio State University

PUBLICATIONS

Wainwright, J., Jiang, S., and Liu, D. 2012 (Accepted). Deforestation and the world-as-representation: the Maya forest of southern Belize. In Land Change Science, and


FIELDS OF STUDY

Major Field: Geography

(Remote Sensing, GIS, Spatial Statistics)
TABLE OF CONTENTS

ABSTRACT ........................................................................................................................................ ii

ACKNOWLEDGMENTS .................................................................................................................. v

VITA ............................................................................................................................................. vii

LIST OF TABLES ................................................................................................................................. xv

LIST OF FIGURES ............................................................................................................................... xvii

CHAPTER 1 INTRODUCTION ........................................................................................................ 1

   1 Background ................................................................................................................................... 1

       1.1 Spatial data quality .................................................................................................................. 1

       1.2 Accuracy assessment in image classification ........................................................................... 2

   2 Research Questions ..................................................................................................................... 4

   3 Objectives ................................................................................................................................... 5

   4 Organization of the Dissertation ................................................................................................. 6

CHAPTER 2 PER-PIXEL CLASSIFICATION CONFIDENCE: A REVIEW ......................... 9

   1 Introduction ................................................................................................................................... 9

   2 Sources of Classification Error ................................................................................................... 9
3 Classification of Methods for Per-Pixel Classification Confidence........................................ 13

3.1 Classification score based method ......................................................................................... 13

3.2 Interpolation based method .................................................................................................. 15

3.3 Regression based method ....................................................................................................... 16

4 Classification Score Based Method.......................................................................................... 17

4.1 Studies using maximum likelihood classification ................................................................. 18

4.2 Studies using non-parametric classifiers ............................................................................... 18

4.3 Studies using multiple classifiers ........................................................................................ 20

5 Interpolation Based Method..................................................................................................... 20

5.1 Local error matrix method ..................................................................................................... 21

5.2 Bootstrap plus interpolation .................................................................................................. 21

5.3 Geostatistical method ........................................................................................................... 22

6 Regression Based Method......................................................................................................... 23

6.1 Map comparison combined with statistical regression ......................................................... 23

6.2 Bootstrap plus statistical regression ..................................................................................... 27

7 Discussion and Conclusions...................................................................................................... 28

7.1 Advantages and limitations .................................................................................................... 28

7.2 Different definitions of per-pixel classification confidence .................................................. 30

7.3 Method validation ................................................................................................................ 33
CHAPTER 3 EVALUATING CLASSIFICATION SCORE BASED METHODS ON PER-PIXEL CLASSIFICATION CONFIDENCE

1 Introduction .................................................................................................................. 39

2 Methodology ................................................................................................................. 41

2.1 Test datasets ............................................................................................................. 42

2.2 Classifier and classification scores .......................................................................... 45

2.3 Probability measures derived from classification scores ........................................ 54

2.4 Approaches to evaluate method performance ....................................................... 55

2.5 Summary of methodology ...................................................................................... 57

3 Results .......................................................................................................................... 58

3.1 Primate probability ................................................................................................. 58

3.2 Margin of victory ...................................................................................................... 67

3.3 Relative entropy ....................................................................................................... 74

3.4 Comparison of three probability measures ........................................................... 81

4 Discussion and Conclusions ....................................................................................... 81

CHAPTER 4 EVALUATING THREE INTERPOLATION BASED METHODS ON PER-PIXEL CLASSIFICATION CONFIDENCE

1 Introduction .................................................................................................................. 83

2 Methodology ................................................................................................................. 84
2.1 Test datasets ................................................................. 85
2.2 Three interpolation based methods ..................................... 85
2.3 Estimated classification confidence vs. true classification confidence .......... 97
2.4 Approaches to evaluating method performance ......................... 98
2.5 Summary of methodology .................................................. 102
3 Results .............................................................................. 103
  3.1 Local error matrix method .................................................. 103
  3.2 Bootstrap method .............................................................. 110
  3.3 Geostatistical method ......................................................... 117
4 Discussion and conclusions .................................................... 120

CHAPTER 5 ESTIMATING PER-PIXEL CLASSIFICATION CONFIDENCE USING CLASSIFICATION SCORES ............................................................. 124

1 Introduction ........................................................................ 124
2 Methodology ...................................................................... 124
  2.1 Test datasets .................................................................. 126
  2.2 Image classification ......................................................... 126
  2.3 Estimate per-pixel classification confidence ......................... 127
  2.4 Method evaluation ........................................................... 130
  2.5 Impact of sample design on PPCC estimates ......................... 133
3 Results

3.1 Maps of classification error and MV

3.2 Estimated PPCC

3.3 Evaluation of PPCC

3.4 Impact of sampling design on PPCC estimates

4 Discussion and Conclusions

CHAPTER 6 CONCLUSIONS AND IMPLICATIONS

1 Conclusions

2 Implications

3 Further studies

References
LIST OF TABLES

Table 2.1 Sources of classification error................................................................. 10
Table 2.2 Summary of classification score based method............................................. 14
Table 2.3 Summary of interpolation based method....................................................... 16
Table 2.4 Summary of regression based method........................................................ 17
Table 2.5 Advantages and limitations of studies on per-pixel classification confidence .29
Table 2.6 Terms and measures for classification confidence ......................................... 32
Table 3.1 Study site and data .................................................................................... 44
Table 3.2 Correlation coefficients (R) of BCQ_q vs. b_q for primate probability .......... 67
Table 3.3 Correlation coefficients (R) of BCQ_q vs. b_q for margin of victory ............. 73
Table 3.4 Correlation coefficients (R) of BCQ_q vs. b_q for relative entropy ............... 80
Table 4.1 A typical error matrix ............................................................................... 93
Table 4.2 Evaluation scheme ..................................................................................... 99
Table 4.3 Willmott's d of estimated and true LCA ..................................................... 107
Table 4.4 Correlation coefficients (R) of BCQ_q vs. b_q for LCA (Q=30) ................. 109
Table 4.5 Willmott's d for estimated and true MR .................................................... 114
Table 4.6 Correlation coefficients (R) of BCQ_q vs. b_q for MR (Q=30) ................. 116
Table 4.7 Correlation coefficients (R) of BCQ_q vs. b_q for CI (Q=30) ..................... 120
Table 4.8 Comparing the results of three interpolation based methods ..................... 121
Table 5.1 Model parameters for BCQ$_q$ vs. $b_q$ based MV of test data ........................................ 138
Table 5.2 Mean and standard deviation for RMSE, Willmott's $d$, and $R^2$ ......................... 146
Table 5.3 $R^2$ for fitting BCQ$_q$ vs. $b_q$ with simple linear model: sample size = 630........ 150
Table 5.4 $R^2$ for fitting BCQ$_q$ vs. $b_q$ with simple linear model: sample size = 1890...... 150
Table 5.5 Two-way ANOVA results for $R^2$ of BCQ$_q$ vs. $b_q$.................................................. 153
Table 5.6 Three-way ANOVA results for $R^2$ of BCQ$_q$ vs. $b_q$.................................................. 153
Table 5.7 $R^2$ for fitting BPC vs. PPCC with the 1:1 reference line: sample size = 630 .... 157
Table 5.8 $R^2$ for fitting BPC vs. PPCC with simple linear model: sample size = 1890 .. 157
Table 5.9 Two-way ANOVA results for $R^2$ of BPC vs. PPCC ................................................. 161
Table 5.10 Three-way ANOVA results for $R^2$ of BPC vs. PPCC ............................................. 161
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>Structure of the dissertation</td>
<td>6</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>Simple diagram for estimating per-pixel classification confidence</td>
<td>13</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Flow chart of the methodology</td>
<td>42</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Datasets and corresponding reference map</td>
<td>43</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Single hidden layer, feed-forward neural network</td>
<td>48</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>AdaBoost.ML: a multiclass logistic version of AdaBoost</td>
<td>52</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>Maps of primate probability based on MLC</td>
<td>59</td>
</tr>
<tr>
<td>Figure 3.6</td>
<td>Maps of primate probability based on NN</td>
<td>59</td>
</tr>
<tr>
<td>Figure 3.7</td>
<td>Maps of primate probability based on SVM</td>
<td>60</td>
</tr>
<tr>
<td>Figure 3.8</td>
<td>Maps of primate probability based on BDT</td>
<td>60</td>
</tr>
<tr>
<td>Figure 3.9</td>
<td>Maps of classification error based on MLC</td>
<td>61</td>
</tr>
<tr>
<td>Figure 3.10</td>
<td>Maps of classification error based on NN</td>
<td>61</td>
</tr>
<tr>
<td>Figure 3.11</td>
<td>Maps of classification error based on SVM</td>
<td>62</td>
</tr>
<tr>
<td>Figure 3.12</td>
<td>Maps of classification error based on BDT</td>
<td>62</td>
</tr>
<tr>
<td>Figure 3.13</td>
<td>Bi-histogram of primate probability (PP) for correct and error pixels</td>
<td>64</td>
</tr>
<tr>
<td>Figure 3.14</td>
<td>BCQ&lt;sub&gt;q&lt;/sub&gt; vs. b&lt;sub&gt;q&lt;/sub&gt; based on primate probability for all the pixels of the whole map</td>
<td>65</td>
</tr>
<tr>
<td>Figure 3.15</td>
<td>BCQ&lt;sub&gt;q&lt;/sub&gt; vs. b&lt;sub&gt;q&lt;/sub&gt; based on primate probability for 2000 test pixels (S&lt;sub&gt;2&lt;/sub&gt;)</td>
<td>66</td>
</tr>
</tbody>
</table>
Figure 3.16 Maps of margin of victory based on MLC .................................................. 68
Figure 3.17 Maps of margin of victory based on NN .................................................... 68
Figure 3.18 Maps of margin of victory based on SVM ................................................. 69
Figure 3.19 Maps of margin of victory based on BDT .................................................. 69
Figure 3.20 Bi-histogram of margin of victory (MV) for correct and error pixels........... 70
Figure 3.21 BCQ_q vs. b_q based on margin of victory for all the pixels of the whole map 72
Figure 3.22 BCQ_q vs. b_q based on margin of victory for 2000 test pixels (S_2)......... 73
Figure 3.23 Maps of relative entropy based on MLC ................................................... 75
Figure 3.24 Maps of relative entropy based on NN ...................................................... 75
Figure 3.25 Maps of relative entropy based on SVM .................................................. 76
Figure 3.26 Maps of relative entropy based on BDT ................................................... 76
Figure 3.27 Bi-histogram of relative entropy (RH) for correct and error pixels .......... 77
Figure 3.28 BCQ_q vs. b_q based on relative entropy for all the pixels of the whole map .. 79
Figure 3.29 BCQ_q vs. b_q based on relative entropy for for 2000 test pixels (S_2)........ 80
Figure 4.1 Flow chart of the methodology ................................................................... 85
Figure 4.2 Grid pixels overlaying with reference map ................................................. 87
Figure 4.3 Example grid pixels and k nearest neighbors ............................................. 87
Figure 4.4 Experimental semivariogram and fitted model for Data 1 classified using MLC ................................................................................................................................. 91
Figure 4.5 Map of estimated LCA from IDW interpolation (MLC) ......................... 104
Figure 4.6 Maps of true LCA (MLC) .......................................................................... 104
Figure 4.7 Indicator maps of classification error (MLC) ........................................... 105
Figure 4.8 Bi-histogram of estimated LCA for correct and error pixels ................. 106
Figure 4.9 Bi-histogram of true and estimated LCA .......................................................... 107
Figure 4.10 Scatter plot of BCQ_q vs. b_q based on LCA for all the pixels.......................... 109
Figure 4.11 Maps of estimated MR from bootstrap method (MLC) ................................. 111
Figure 4.12 Maps of true MR from bootstrap method (MLC) ........................................ 111
Figure 4.13 Indicator maps of true classification error from bootstrap method (MLC) .... 112
Figure 4.14 Bi-histogram of MR for correct and error pixels......................................... 113
Figure 4.15 Bi-histogram of true and estimated MR ....................................................... 114
Figure 4.16 Scatter plot of BCQ_q vs. b_q based on estimated MR ................................. 116
Figure 4.17 Map of estimated CI (MLC) .................................................................... 117
Figure 4.18 Bi-histogram of CI ...................................................................................... 118
Figure 4.19 Scatter plot of BCQ_q vs. b_q based on CI .................................................. 119
Figure 5.1 Flow chart of the methodology .................................................................... 125
Figure 5.2 Indicator maps of classification error (MLC) .................................................. 136
Figure 5.3 Maps of MV based on MLC ........................................................................ 136
Figure 5.4 Scatter plot of BCQ_q vs. b_q based on MV of test data .................................. 138
Figure 5.5 Predicted PPCC using MV based on MLC .................................................... 140
Figure 5.6 PPCC by class for Data 1 using MLC .............................................................. 140
Figure 5.7 PPCC by class for Data 1 using NN ................................................................. 141
Figure 5.8 PPCC by class for Data 1 using SVM ............................................................. 141
Figure 5.9 PPCC by class for Data 1 using BDT ............................................................. 142
Figure 5.10 PPCC by class for Data 2 using MLC ........................................................... 142
Figure 5.11 PPCC by class for Data 3 using MLC ........................................................... 143
Figure 5.12 Bi-histogram of PPCC for correct and error pixels ...................................... 144
Figure 5.13 Scatter plot of BPC vs. PPCC ................................................................. 145
Figure 5.14 Scatter plot of BCQ_q vs. b_q with MLC for 630 test pixels: SRS .............. 147
Figure 5.15 Scatter plot of BCQ_q vs. b_q with MLC for 630 test pixels: STS ............. 148
Figure 5.16 Scatter plot of BCQ_q vs. b_q with MLC for 630 test pixels: SYS ............. 148
Figure 5.17 Scatter plot of BCQ_q vs. b_q with MLC for 630 test pixels: CRS ............ 149
Figure 5.18 Impact of sample design on the relationship between BCQ_q and b_q ....... 151
Figure 5.19 Scatter plot of BPC vs. PPCC with MLC for 630 test pixels: SRS .......... 154
Figure 5.20 Scatter plot of BPC vs. PPCC with MLC for 630 test pixels: STS .......... 155
Figure 5.21 Scatter plot of BPC vs. PPCC with MLC for 630 test pixels: SYS .......... 155
Figure 5.22 Scatter plot of BPC vs. PPCC with MLC for 630 test pixels: CRS ........... 156
Figure 5.23 Impact of sample design on the relationship between BPC vs. PPCC ...... 158
Figure 5.24 Impact of sample design on the relationship between BPC vs. PPCC for MLC, NN, and SVM ........................................................................... 159
CHAPTER 1 INTRODUCTION

1 Background

1.1 Spatial data quality

Spatial data quality is an important topic in geographic information sciences and remote sensing, which has drawn attention from academic community, government agencies, and industry (Devillers et al. 2010). Studies on the quality of spatial data went back as early as to the work in the 1970s. It was pointed out that errors and uncertainties are inherent to spatial data and can propagate to further spatial analysis (Goodchild 1978). With the development of GIS and remote sensing, spatial data quality has been an continuing hot topic in the research community during the last three decades. Papers and books on spatial data quality have been regularly published (see e.g., Beard 1989; Bédard 1987; Burrough et al. 1996; Burrough 1992; Buttenfield 1993; Chrisman 1991; Congalton and Green 2009; Delavar and Devillers 2010; Devillers et al. 2007; Devillers and Goodchild 2009; Devillers and Jeansoulin 2006; Fisher 1999; Foody and Atkinson 2002; Goodchild 1991; Goodchild and Jeansoulin 1999; Goodchild and Gopal 1989; Goodchild and Li 2012; Guptill and Morrison 1995; Shi 2009; Shi et al. 2002; Stein et al. 2008; Zhang and Goodchild 2002). Spatial data quality become more and more important with the increasing use of location systems (e.g., GPS, mobile GIS, etc.) and high spatial/spectral remote sensing images (Delavar and Devillers 2010).
Although great progress has been made on the spatial quality of interval and ratio data, the spatial uncertainty of nominal and ordinal data remains problematic (Goodchild 2003). Land use land cover is one of the most important nominal data, which has broad impacts on our environment. Land use land cover change (LULCC) affects the environment by changing carbon pools and flux (Dixon et al. 1994; Guo and Gifford 2002; Kaplan et al. 2012). It is found that LULCC is one of the key factors influencing global warming (Vitousek 1994). Therefore, land cover change is one important variable used to simulate future climates (Feddema et al. 2005). Besides the impact on climate, LULCC also has significant effects on biodiversity (Chapin et al. 2000; Sala et al. 2000). The significance of LULCC as an environmental factor calls for studies on the spatial data quality in LULCC.

Remote sensing image classification is the most common source for LULCC. As a type of nominal data, LULCC data are usually constructed using two types of data models: vector and raster (Goodchild et al. 1992). The vector model delineates homogeneous areas with lines or polygons while the raster model identifies class label for each cell of a regular tessellation. This study focuses on the raster class map which is usually derived from remote sensing image classification. Therefore, the question of spatial data quality in LULCC turns to the accuracy assessment in image classification.

1.2 Accuracy assessment in image classification

Image classification usually provides two types of information: class label assignment and classification uncertainty (Atkinson and Foody 2002). Accordingly, two types of models are required: Model I is the classification algorithm which determines the class labels assigned to pixels; Model II is the accuracy assessment and gives information
on the uncertainty of the class assignment. Traditionally, Model II is in the form of an error matrix and thus global accuracy indices such as overall accuracy, user's accuracy, producer's accuracy, and kappa are derived for the whole image (Congalton et al. 1983; Rosenfield and Fitzpatrick-lins 1986; Story and Congalton 1986). The error matrix approach is commonly practiced for accuracy assessment as it generates essential measures for overall map quality (Congalton 1991; Congalton and Green 2009; Foody 2002; Jansen and van der Wel 1994).

Reporting classification confidence at pixel level is an emerging challenge to accuracy assessment in remote sensing image classification. Campbell (1981) is one of the earliest studies recognizing the spatial variation of classification confidence. It is now widely accepted that classification confidence is neither uniformly nor randomly distributed across the classification map (Congalton 1988b; Foody 2002; Steele et al. 1998). Factors contributing to the spatial variability of classification confidence include ground features (such as topography and elevation), land cover type and heterogeneity, patch size, and sample design (Burnicki 2011; Congalton 1988b; Smith et al. 2002; van Oort et al. 2004; Yu et al. 2008). Irrespective of its source, spatial pattern of classification confidence will propagate to further applications using error-infected maps, and ignorance of the spatial variability of confidence may have negative impact on decision making (McIver and Friedl 2001; Steele et al. 1998; van Oort et al. 2004). The traditional approach for accuracy assessment is limited in that error matrix and derived indices provide no information on the classification confidence at pixel level. Therefore, it is important to develop methods to identify the spatial distribution of classification confidence and integrate it into analysis afterwards.
Previous studies have tried various methods to estimate per-pixel classification confidence. Here is a short list of previous studies on per-pixel classification confidence whereas a detailed literature review will be presented in Chapter 2. Foody et al. (1992) and Goodchild et al. (1992) find that classification scores output from classifiers may be used as indicators for classification confidence at pixel level. Maselli et al. (1994) represent classification confidence using relative probability entropy based on probability output from classifiers. Steele et al. (1998) present a method to estimate misclassification probability maps using bootstrap sampling and kriging. McIver and Friedl (2001) classify image with boosted decision tree and examine the relationship between classification accuracy/error with three measures. Liu et al. (2004) propose a hybrid classifier combining ARTMAP neural network and decision tree to provide per-pixel classification confidence. Foody (2005) obtained local classification accuracy by partitioning sample data into subregions and generating an individual error matrix for each subregion. The classification accuracy in each subregion was then interpolated to pixel-level across the whole map. Burnicki (2011) adopted a generalized additive model (GAM) to examine the relationship between classification error and landscape composition and structure. Comber et al. (2012) use geographically weighted regression to model the spatial variation of classification uncertainty.

2 Research Questions

Although a lot of studies have been devoted to per-pixel classification confidence during the past two decades, several limitations hinder the wider practice of per-pixel classification confidence. First, there is no comprehensive review in this field. Due to the lack of literature review, we do not know the status of per-pixel classification confidence.
Second, previous studies have not been rigorously validated with empirical evidence. Although some studies have validated their results using certain approaches, they are far from adequate. Many studies usually report observations based on limited number of sample data, the estimated per-pixel classification confidence has not been comprehensively validated with true per-pixel classification confidence. Third, there are not many studies comparing the performance of different methods. Partly due to the fact that each single method is not well validated with reference data, comparisons between different methods are not available. Fourth, there is no consensus method on estimating classification confidence at pixel level. While new methods are increasingly developed, no method is accepted as the benchmark for identifying per-pixel classification confidence.

With the increasing need for a sound method to predict classification confidence at pixel level, it is imperative to address the above four closely related limitations. This dissertation answers the following three questions.

(1) What is the status of per-pixel classification confidence?

(2) How efficient are the previous methods in estimating per-pixel classification confidence?

(3) What's the benchmark method to estimate per-pixel classification confidence?

3 Objectives

The overall objective of my dissertation is to develop a method to characterize per-pixel classification confidence. In order to realize this overarching goal, I aim to achieve the following three objectives in this dissertation, which answers the research questions proposed in Section 2.
(1) To provide a comprehensive review of previous methods on per-pixel classification confidence;

(2) To conduct a rigorous evaluation of previous methods;

(3) To develop a method to estimate per-pixel classification confidence that is comparable across different classifiers.

4 Organization of the Dissertation

This dissertation consists of six chapters. According to the diagram in Figure 1.1, the dissertation is organized as follows:

Figure 1.1 Structure of the dissertation
Chapter 1 briefly introduces spatial data quality, accuracy assessment in remote sensing image classification; identifies research questions; proposes research objectives; and describes organization of the dissertation.

Chapter 2 reviews previous studies related to per-pixel classification confidence. In this chapter I provide a comprehensive review of previous studies in estimating classification confidence at pixel level. I classify previous methods into three categories. Literature on methods in each category is reviewed. The advantages and limitations of previous methods are discussed. I point out that it is imperative to conduct a rigorous evaluation of previous methods, which is the work of Chapters 3 and 4.

Chapter 3 evaluates classification score based method (CSBM). First, I give a brief development of the algorithms to derive classification scores from four commonly used classifiers: Maximum Likelihood Classification (MLC), Neural Network (NN), Support Vector Machines (SVM), and Boosted Decision Tree (BDT). Second, I present the procedures to calculate three probability measures. Third I evaluate the usefulness of probability measures as classification confidence at pixel level.

Chapter 4 evaluates three interpolation based methods (IBM): local error matrix method, bootstrap method, and geostatistical method. First, I introduce the procedures to estimate per-pixel classification confidence using three methods. Then, I present approaches to evaluate these methods. Third, results for the evaluation are provided afterwards. The results of Chapters 3 and 4 suggest that a new
approach should be developed to estimate per-pixel classification confidence based on classification scores.

**Chapter 5** develops and evaluates a new method to estimate per-pixel classification confidence (PPCC) using classification scores. First, I present a new method to estimate per-pixel classification confidence using classification scores. Then I evaluate the method in three aspects: (1) statistical relationship of binned PPCC vs. PPCC, (2) stability of PPCC estimates, and (3) impact of sample design on PPCC estimates.

**Chapter 6** concludes the dissertation and discusses implications for future work.
CHAPTER 2 PER-PIXEL CLASSIFICATION CONFIDENCE: A REVIEW

1 Introduction

Per-pixel classification confidence is an important topic in the remote sensing community. It has attracted continuous attention in the past two decades. Although various methods have been proposed to estimate classification confidence at pixel level, there is no comprehensive review in this field. This chapter aims to fill this gap.

The remainder of this chapter is organized as follows. First, sources of classification error especially the spatial distribution of error are reviewed in Section 2. Second, a classification of previous methods is presented in Section 3. Next, three categories of methods are reviewed in Section 4, 5, and 6 respectively. Section 7 discusses the current status of per-pixel classification confidence and concludes the chapter. The review in this chapter calls for further studies on the evaluation of major methods, which is the work of Chapter 3 and 4.

2 Sources of Classification Error

Previous work on the sources of classification error can be divided into two groups: (1) error analysis for a single-date classified map and (2) classification error in land-cover change detection. Most of the previous studies belong to the first group, and there are
only a limited number of papers in the second group (Burnicki 2011). Table 2.1 shows the main studies on classification error in the literature.

Table 2.1 Sources of classification error

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Method</th>
<th>Example</th>
<th>Error source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-date map</td>
<td>Visual comparison</td>
<td>Congalton (1988a, b)</td>
<td>sensor systems, platforms, ground control, radiometric rectification, geometric rectification, classification algorithms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lunetta et al. (1991)</td>
<td>ground features (such as topography), land cover types, and sampling techniques</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friedl et al. (2000)</td>
<td>Sampling design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Steele et al. (1998)</td>
<td>terrain, landscape complexity, and land use patterns</td>
</tr>
<tr>
<td></td>
<td>Log linear model</td>
<td>Foody and Arora (1997)</td>
<td>dimensionality of the dataset, training and testing set characteristics</td>
</tr>
<tr>
<td></td>
<td>Linear mixed model</td>
<td>Moisen et al. (2000)</td>
<td>stratum, slope, and local heterogeneity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yu et al. (2008)</td>
<td>elevation, sample object size, sample object reliability, sample object density, and sample spatial composition</td>
</tr>
<tr>
<td></td>
<td>Logistic regression</td>
<td>Smith et al. (2002)</td>
<td>land-cover heterogeneity and patch size</td>
</tr>
<tr>
<td>Land cover change</td>
<td>Mixed regression</td>
<td>Carmel (2004)</td>
<td>topographic structure, slope, and vegetation cover</td>
</tr>
<tr>
<td></td>
<td>Generalized additive model</td>
<td>Burnicki (2011)</td>
<td>landscape composition and structure</td>
</tr>
</tbody>
</table>
In the literature, both qualitative and quantitative methods are used to study the classification error for single-date classified map.

For a qualitative approach, researchers visually compare pattern similarity and summarize accuracy affected by different factors. Lunetta et al. (1991) provide an interesting overview on error accumulation in the whole process of remote sensing and GIS data analysis. Among the factors and procedures discussed in Lunetta et al. (1991), the first three aspects, i.e., data acquisition, data processing, and data analysis, all contribute to classification confidence. In other words, factors that impact classification accuracy include sensor systems, platforms, ground control, scene consideration, radiometric rectification, geometric rectification, classification algorithms. In two earlier studies, Congalton (1988a, b) finds that the patterns of classification error is influenced by ground features (such as topography), land cover types, and sampling techniques. The impact of sampling design on classification accuracy has also been verified by other researchers such as Dicks and Lo (1990), Gong and Howarth (1990), Stehman (1992), and Friedl et al. (2000), as well as the review paper of Foody (2002). In one study on the misclassification probabilities for a land cover map, Steele et al. (1998) reveal the impact of terrain, landscape complexity, and land use patterns on the spatial distribution of classification error.

For quantitative studies, several researchers have used statistical models to examine the relationship between classification error and potential factors. Models used in the literature include log-linear model (Foody and Arora 1997), linear mixed model (Moisen et al. 2000; Yu et al. 2008), and logistic regression model (Smith et al. 2003; Smith et al. 2002; van Oort et al. 2004). For example, Foody and Arora (1997) use a log-linear model
to evaluate the impact of factors on classification accuracy. They find that variations in the dimensionality of the dataset, as well as the training and testing set characteristics have a significant effect on classification accuracy. Moisen et al. (2000) apply a generalized linear mixed model to account for spatial autocorrelation and identify strong relationship between map error and stratum, slope, and local heterogeneity. Yu et al. (2008) also use a linear mixed model but in an object-based classification case. Their results show that elevation, sample object size, sample object reliability, sample object density, and sample spatial composition significantly influence the object-based classification confidence. Smith et al. (2003; 2002) and van Oort et al. (2004) use logistic regression models to quantify the relationship between classification accuracy and landscape variables. They find that land-cover heterogeneity and patch size are significant variables affecting the location of classification error.

As Burnicki (2011) points out, there are considerably less studies on the spatial distribution of classification error in land cover change. Carmel (2004) uses a mixed regression model and finds that topographic structure, slope, and vegetation cover have significant effects on classification error. van Oort (2007) finds that land cover class boundary over time is an important factor affecting temporally correlated error pattern. A recent study by Burnicki (2011) using generalized additive model reveals that landscape composition and structure combined with classification error in single-date maps are significant factors for the spatial distribution of error in land cover change detection.

In summary, key factors impacting the spatial distribution of classification error can be summarized into the following three categories:

(1) System factors: sensor, platform, radiometric and geometric rectification;
(2) Object factors: topography, land use land cover pattern, landscape complexity;

(3) Interpretation factors: classification algorithm, sample design.

3 Classification of Methods for Per-Pixel Classification Confidence

The process for estimating per-pixel classification confidence is composed of three parts: input, model, and output (Figure 2.1). The output is classification confidence at pixel level. There are two types of variables for classification confidence: (1) binary variable indicating a pixel being correctly or incorrectly classified; (2) continuous probability of classification confidence indicating the probability of each pixel being correctly classified. Depending on the input and modeling techniques used, methods for per-pixel classification confidence can be divided into three categories: classification score based method, interpolation based method, and regression based method. The categories of reviewed methods are shown in Table 2.2 - Table 2.4. A simple discussion of three categories of methods is presented in this section, while detailed reviews follow in Sections 4-6.

![Simple diagram for estimating per-pixel classification confidence](image)

Figure 2.1 Simple diagram for estimating per-pixel classification confidence

3.1 Classification score based method

Classification score based method (CSBM) derives per-pixel classification confidence using classification scores from classifiers. Table 2.2 shows main studies on estimating per-pixel classification confidence using classification score based method.
### Table 2.2 Summary of classification score based method

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Input</th>
<th>Output</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum likelihood classification (MLC)</td>
<td>Posterior probabilities</td>
<td>Posterior probabilities</td>
<td>Foody et al. (1992)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Probability entropy</td>
<td>Maselli et al. (1994)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Absolute accuracy, relative accuracy, mixture level, and incompleteness</td>
<td>Shi (1994, 2010)</td>
</tr>
<tr>
<td></td>
<td>Distance to cluster centroids</td>
<td>Standardized distance to cluster centroids and related measures</td>
<td>Mitchell et al. (2008)</td>
</tr>
<tr>
<td>Neural network (NN)</td>
<td>Posterior probabilities</td>
<td>Normalized maximum possibility</td>
<td>Gong et al. (1996)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Activation level</td>
<td>Foody (2000)</td>
</tr>
<tr>
<td>Boosted decision tree (BDT)</td>
<td>Posterior probabilities</td>
<td>Posterior probabilities, margin of victory, entropy</td>
<td>McIver and Friedl (2001)</td>
</tr>
<tr>
<td>K nearest neighbor (k-NN)</td>
<td>Neighborhood information</td>
<td>A series of measures based on distance and neighborhood measures</td>
<td>Delany et al. (2005)</td>
</tr>
<tr>
<td>NN and decision tree (DT)</td>
<td>Posterior probabilities</td>
<td>Posterior probabilities</td>
<td>Liu et al. (2004)</td>
</tr>
<tr>
<td>MLC and NN</td>
<td>Posterior probabilities, activation level</td>
<td>Average posterior probabilities and activation level, entropy</td>
<td>Brown et al. (2009)</td>
</tr>
</tbody>
</table>

The procedures of classification score based methods are summarized as follows.

1. **Input**: Obtaining classification scores at each pixel.

   The most widely used classification scores are the (posterior) probabilities of class membership, also called probability vectors, from different classifiers. Besides posterior
probabilities, other measures used in the literature include activation level, distance to neighbors or cluster centroids. The most commonly used classifier is maximum likelihood classifier (MLC). Besides MLC, other classifiers frequently used include neural network (NN), decision tree (DT), boosted decision tree (BDT), $k$-nearest neighbor ($k$-NN), etc.

(2) Model: constructing models for per-pixel classification confidence.

(3) Output: applying models to produce classification confidence for each pixel.

Usually a simple transformation of the classification scores is applied to generate per-pixel classification confidence. The resultant measures include posterior probabilities, weighted posterior probabilities, entropy, normalized maximum possibility, margin of victory, binned average of posterior probabilities and activation level, absolute accuracy, relative accuracy, mixture level, and incompleteness, standardized distance to cluster centroids and related measures, and other measures based on distance and neighborhood measures, etc..

### 3.2 Interpolation based method

Interpolation based method (IBM) models classification confidence at pixel level by interpolating estimation at sample pixels to the whole image. Table 2.3 shows main literature in this field.

The process of interpolation based method is summarized as follows.

(1) Input: Classification confidence at sample pixels.

Classification confidence is generated using local error matrix, bootstrap classification or geostatistical method.

(2) Model: Interpolation techniques.
(3) Output: Interpolating classification confidence at sample pixels to all the pixels using the interpolation models from (2).

<table>
<thead>
<tr>
<th>Methods</th>
<th>Models for generating output data</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local error matrix</td>
<td>Inverse distance squared interpolation</td>
<td>Foody (2005)</td>
</tr>
<tr>
<td>Bootstrap classification</td>
<td>Kriging</td>
<td>Steele et al. (1998)</td>
</tr>
<tr>
<td>Geostatistical method</td>
<td>Simple indicator kriging with varying local mean (SIK)</td>
<td>Kyriakidis and Dungan (2001)</td>
</tr>
</tbody>
</table>

Two types of interpolation models are commonly used in the literature: inverse distance weighted interpolation (IDW) and kriging. For IDW, the weights are calculated based on neighboring pixels. For kriging, semivariogram model is fitted based on sample pixels, which is then used to compute the interpolation weights.

### 3.3 Regression based method

RBM analyzes the factors impacting classification error and then uses those significant factors to predict the spatial distribution of classification confidence. Table 2.4 shows main studies using regression based method. The process of regression based method is summarized as follows.

(1) Input: classification confidence at sample pixels.

There are two types of variables for input classification confidence: binary variable and continuous classification confidence. The binary variable is represented as indicator
variable indicating a pixel as correctly classified/incorrectly classified. The continuous classification confidence variable is represented as the probabilities of a pixel being correctly classified.

(2) Model: fitting regression models between input classification confidence and covariates at sample locations.

(3) Output: applying models to predicate classification confidence for all the pixels using significant factors.

Models using in the regression include: linear mixed model, logistic model, generalized additive model, and geographically weighted regression.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Models for generating output data</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Map comparison</strong></td>
<td>Linear mixed model</td>
<td>Moisen et al. (2000)</td>
</tr>
<tr>
<td></td>
<td>Logistic model</td>
<td>Smith et al. (2002)</td>
</tr>
<tr>
<td></td>
<td>Generalized additive model</td>
<td>Burnicki (2011)</td>
</tr>
<tr>
<td></td>
<td>Geographically weighted regression</td>
<td>Comber et al. (2012)</td>
</tr>
<tr>
<td><strong>Bootstrap</strong></td>
<td>Linear mixed model</td>
<td>Yu et al. (2008)</td>
</tr>
</tbody>
</table>

4 Classification Score Based Method

In remote sensing image classification, classifiers usually assign class to pixels based on the classification scores. Foody et al. (1992) and Goodchild et al. (1992) propose that classification scores may be used to characterize per-pixel classification confidence. During the past two decades, the classification score based method has
attracted attention from researchers both in geographic information science and remote sensing communities. The following review is based on the summary in Table 2.2.

4.1 Studies using maximum likelihood classification

In early studies, many users used the posterior probabilities from maximum likelihood classification as an input to derive measures for per-pixel classification confidence. In Foody et al. (1992)’s study, posterior probabilities from maximum likelihood classification are directly used to represent per-pixel classification confidence. Later, Maselli et al. (1994) derive a new measure called probability entropy and use it to represent classification confidence at pixel level. The relative entropy is calculated using posterior probabilities from maximum likelihood classification. Besides, probability entropy, other measures are also developed. For example, Shi (1994, 2010) derived four measures based on posterior probabilities: absolute accuracy, relative accuracy, mixture level, and incompleteness. Canters (1997) derived weighted posterior probabilities to map per-pixel classification confidence.

Besides posterior probabilities, distance to cluster centroid is also used to indicate classification confidence. For example, Mitchell et al. (2008) use distance to second cluster as a measure for classification confidence.

4.2 Studies using non-parametric classifiers

Maximum likelihood classifier is a parametric method which assumes a Gaussian distribution of the data. However, this is not necessarily the case for real data. Therefore, non-parametric classifiers may provide more accurate solution to image classification. In this regard, several researchers have used neural network, boosted decision tree and k-nearest neighbor to derive per-pixel classification confidence.
Gong et al. (1996) is one of the early studies using neural network to estimate the spatial distribution of classification confidence. It is believed that classification confidence is related to the maximum posterior probability, $p_k$, and $1 - p_k$ represent the uncertainty of a classified pixel. In order to consider the second largest posterior probability, Gong et al. (1996) introduce a measure called normalized maximum possibility

$$p^* = \frac{p_k}{\sum_{i=1}^{K} p_i}$$  \hspace{1cm} (2.1)

where $K$ is the number of classes. The uncertainty factor is then defined as

$$1 - \frac{p_k^2}{\sum_{i=1}^{K} p_i}$$  \hspace{1cm} (2.2)

Equations (2.1) - (2.2) consider the distribution of posterior probabilities in different classes and thus may be more accurate for characterizing per-pixel classification confidence. Foody (2000) is another example using neural network to examine the per-pixel classification confidence. Similar to previous studies using maximum likelihood classification, Foody (2000) calculate entropy $H$ using the activation level from neural network. This entropy $H$ is then used to map spatial distribution of classification confidence.

Boosted decision tree is a non-parametric classifier commonly used in machine learning. It has also been introduced into the remote sensing literature. For example, McIver and Friedl (2001) use boosted decision tree to classify image and derive measures for per-pixel classification confidence. Besides the direct use of posterior probability and the commonly used entropy measure, McIver and Friedl (2001) also introduced a new
measure, called margin of victory, which is the difference in the posterior probabilities of the two most probable classes. Entropy, margin of victory, and the normalized maximum possibility measure of Gong et al. (1996) are similar in that they all consider the influence of the distribution of posterior probabilities among classes.

Delany et al. (2005) find that confidence measures based on numeric scores from Naive Bayes, SVM are not consistent at predicting confidence in the spam filtering domain. Therefore, they propose a method to estimate classification confidence using a series of measures based on distance and neighborhood information: Average Nearest Unlike Neighbour Index (Avg NUN Index), Similarity Ratio, Similarity Ratio within K, Sum of NN Similarities, and Average NN Similarity.

4.3 Studies using multiple classifiers

With the successful application of different classifiers in the study of per-pixel classification confidence, some researchers have also proposed methods by combining multiple classifiers. For example, Liu et al. (2004) propose a hybrid classifier method by combining ARTMAP neural network and decision tree to produce per-pixel classification confidence. Some researchers also compare the performance of different classifiers in estimating per-pixel classification confidence. Brown et al. (2009) compare the performance of maximum likelihood classifier and two types of neural network classifiers in predicting thematic accuracy and classification uncertainty.

5 Interpolation Based Method

For interpolation based method, a interpolation model is constructed based on sample data and then applied to the whole map to generate classification confidence at pixel level. Two types of interpolation models are used in the literature: inverse distance
weighted interpolation and kriging interpolation. The following review is based on the summary in Table 2.3.

5.1 Local error matrix method

Foody (2005) propose a method to combine local error matrix and interpolation to predict per-pixel classification confidence. I call this method local error matrix method. A simple summary of the procedures is given bellow, while the detailed steps will be introduced in Section 2.2.1 of Chapter 4.

First, classify the image in the usual way and obtain a class map.

Second, create a random selection of \( N \) pixels as the test set. Select the \( k \) test pixels closest to the location of the intersection of a \( m \times n \) pixel grid. Local error matrix is created based on these \( k \) test pixels and local classification accuracy is derived.

Third, the local classification accuracy is then interpolated to the whole map using inverse distance squared interpolation.

5.2 Bootstrap plus interpolation

Some researchers have used bootstrap to generate classification confidence at sample locations. The bootstrapped classification confidence can then be interpolated to the whole map to predict per-pixel classification confidence. Here I briefly introduce the work from Steele et al. (1998), while the detailed procedures will be introduced in Section 2.2.2 of Chapter 4. The procedures include two main steps.

Step 1: Bootstrap image classification. Using bootstrap sampling, each sample pixel is classified with \( B \) times. Compare the \( B \) classification results with reference data and record the frequency of each pixel being incorrectly classified. Define misclassification probability \( (p) \) as the frequency of misclassification divided by \( B \).
Step 2: model misclassification probability with semivariogram and interpolate it to the whole map using kriging. To get a better view of the results, Steele et al. (1998) also create a contour map based on the interpolation map.

5.3 Geostatistical method

Kyriakidis and Dungan (2001) provide a study by combining error matrix with kriging to predict per-pixel classification confidence. A simple summary of their method is given below, while the detailed procedures will be given in Section 2.2.3 of Chapter 4.

Step 1, probability and indicator coding of class labels is constructed using global error matrix and test data.

Step 2, residuals are modeled based on the class probabilities and indicator data from step 1 above.

Step 3, the residuals are modeled with semivariogram and interpolated to the whole map using simple kriging. The predicted residuals is then added with class probabilities from step 1 to produce final interpolated class probabilities $p(u;k)$. $p(u;k)$ represents the probability of pixel $u$ belonging to class $k$.

Step 4, local index is constructed based on $p(u;k)$ as

$$c(u) = [1 - p^m(u)] \left( \frac{K}{K-1} \right),$$

(2.3)

where $p^m(u) = \max\{p(u;k), k = 1,\ldots,K\}$, $K$ is the number of classes, and $K/(K-1)$ is a standardization factor.

The central idea of Kyriakidis and Dungan (2001) is to interpret elements in the error matrix as the expected mean for the probability of class $k$ observed for pixel $u$. This
information is then integrated with indicator class labels to model the spatial variation of class probabilities in the whole map.

6 Regression Based Method

Statistical regression was originally devised to analyze the factors impacting classification error. It can also be used to predict the spatial distribution of classification confidence. For regression based method, a regression model is fitted between classification confidence and covariates. The fitted model is then used to predict classification confidence for all the pixels using the significant factors. As discussed in Section 3.3, there are two types of sample data as input variables: (1) indicator values for classification error; (2) continuous probability of classification confidence. The following review is based on the summary in Table 2.4.

6.1 Map comparison combined with statistical regression

For this category of methods, indicator dependent variable is generated by comparing the class map with the reference map. The regression model used in the literature can be divided into two categories: global model and local model. The global model is most familiar to us, where the whole map is fitted with one mathematical model with coefficients the same for all the pixels. The local model approach fits mathematical models locally tuned to the pixel locations. One commonly used local model approach is the geographically weighted regression (Brunsdon et al. 1996; Fotheringham et al. 2002).

6.1.1 Global regression model

Most of the studies using regression based method belong to this category. Statistical models used include linear mixed regression (Carmel 2004; Moisen et al. 2000), logistic regression (Smith et al. 2003; Smith et al. 2002; van Oort et al. 2004),
generalized additive model (Burnicki 2011). The dependent variable takes the form of an indicator variable: correctly classified or incorrectly classified. This indicator variable is created by comparing the class map with the reference map. The explanatory variables are from the significant factors contributing to classification confidence, which usually include topography, land use land cover pattern, landscape complexity, and sampling design.

Moisen et al. (2000) is one of the early studies that use statistical regression to model the spatial distribution of classification error. In their study, they use a generalized linear mixed model to characterize the relationship between classification error and topographical and landscape heterogeneity. The specification of the model is shown in Equation (2.4) - (2.6).

\[ y = \mu + \varepsilon. \] (2.4)

where \( \mu \) is linked using a logit function

\[ g(\mu) = \log \left( \frac{\mu}{1-\mu} \right) + \varepsilon, \] (2.5)

and model

\[ g(\mu) = X\beta + Z\nu. \] (2.6)

\( X \) and \( Z \) are fixed effect and random effect covariates. Nine fixed effects variables are considered: elevation, slope, aspect, richness, evenness, diversity 1, diversity 2, minimum distance to a different map cover, and road stratum. The 96 sample pixels are randomly selected from 15 quadrangles. The quadrangle maps are randomly selected for subsampling and thus acted as random effect variable. The results show that there is a
strong relationship between classification error and stratum, slope, and local heterogeneity.

Similarly, Carmel (2004) also uses a mixed linear model to study the classification error. Carmel finds that topographic structure, vegetation cover, aspects, slope has strong correlation with classification error.

Several researchers use logistical regression to reveal the spatial pattern of classification error (Smith et al. 2003; Smith et al. 2002; van Oort et al. 2004). The model is specified as follows,

$$
\log \left( \frac{p}{1-p} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k,
$$

(2.7)

where $p$ is the probability of correct classification, $x_1$ through $x_k$ are explanatory variables. Smith et al. (2003; 2002) use patch size, heterogeneity, and regions as explanatory variables and find strong correlation between classification error and patch size and heterogeneity. van Oort et al. (2004) extend the work of Smith et al. (2003; 2002) by adding landscape indices and thus include four categories of factors: (1) land cover class, (2) focal heterogeneity and homogeneity, (3) patch size, (4) landscape heterogeneity, dominance, entropy, and contagion. They find that focal heterogeneity, landscape heterogeneity, and patch size have significant influence on per cell classification accuracy.

In a recent study on the misclassification in land cover change, Burnicki (2011) adopted a generalized additive model (GAM) to examine the relationship between classification error and landscape composition and structure. The model is designed as follows
where \( y \) is the response variable, i.e., indicator variable indication classification error, \( X_i \) are the explanatory variables, \( f_i \) are the smoothed functions for each explanatory variable, and \( p \) is the number of explanatory variables. Cubic spline functions are chosen as smooth functions for \( f_i \). Four sets of explanatory variables, or predictors, are selected to fit the model: (1) temporal error information; (2) landscape heterogeneity, land cover patch size, etc; (3) locational information; (4) land-cover types. Strong relationship has been identified between classification error and landscape composition and structure.

In summary, map comparison plus global regression is a straightforward approach to tackle factors impacting classification error. Once the relationship between classification error and explanatory variables is established, we can use the model to predict classification error in the whole map. The model is fitted based on sample data which are assumed to be representative of the whole map. The difficulty lies in the selection of predictors, which requires extra efforts in variable construction and data collection.

### 6.1.2 Geographically weighted regression

Different from the above studies in Section 6.1.1, Comber et al. (2012) use logistic geographically weighted regression to predict spatial classification accuracy for each class.

First, define indicator variables for the reference and classified class label of pixel \( u \) respectively. Let \( s(u) \) be the reference class label of pixel \( u \), and \( x(u) \) the classified class label. Define indicator variable for reference data as
and indicator variable for classified map as

\[ I(u;k) = \begin{cases} 
1, & \text{if } s(u) = k \\
0, & \text{if not } k = 1,2,\ldots,K.
\end{cases} \quad (2.9) \]

where \( k \) is the class label, \( K \) is the total number of classes.

Second, define the logit function as

\[ \text{logit}(Q) = \frac{e^Q}{1 + e^Q} \quad (2.11) \]

where \( Q \) can be any value.

Third, the logistic geographically weighted regression is formulated as follows

\[ pr[I(u;k)] = \text{logit}[b_0(u) + b_1(u)J(u;k)] \quad (2.12) \]

where \( pr[I(u;k)] \) is the probability that the reference class for pixel at location \( u \) is \( k \) or not \( k \). The coefficients \( b_0 \) and \( b_1 \) is obtained through geographically weighted regression.

The probabilities of correctly assigning a pixel to each class is then estimated as

\[ pr[I(u;k) = J(u;k)] = pr[I(u;k) = 1|J(u;k) = 1] \cdot pr[J(u;k) = 1] + pr[I(u;k) = 0|J(u;k) = 0] \cdot pr[J(u;k) = 0] \quad (2.13) \]

Since geographically weighted regression is used, \( pr[I(u;k) = J(u;k)] \) varies with pixel location \( u \).

6.2 Bootstrap plus statistical regression

As I reviewed in Section 5.2, bootstrap can be used to generate classification confidence at sample locations. Using this classification confidence as input, statistical
regression is further applied to predict classification confidence for all the pixels. One typical work is from Yu et al. (2008). The method includes two steps as follows:

Step 1: Bootstrap image classification, same as that in Section 5.2. Using bootstrap sampling, each sample pixel is classified for $B$ times. Compare the $B$ classification results with test data and record the frequency of each pixel being correctly classified. Define classification uncertainty ($c$) for each sample pixel as the ratio of the number of times being correctly classified to the total number of times being classified.

Yu et al. (2008)'s classification uncertainty, $c$, is complementary to Steele et al. (1998)'s misclassification probability, $p$, i.e., $c = 1 - p$.

Step 2: Statistical regression similar to those methods reviewed in Section 6.1. Yu et al. (2008) use linear mixed regression to model the relationship between classification uncertainty and predictors. They find strong relationships between classification uncertainty and elevation, sample object size, sample object reliability, sample object density, and sample spatial composition.

7 Discussion and Conclusions

7.1 Advantages and limitations

A summary of the general advantages and limitations of the existing methods on per-pixel classification confidence is shown in Table 2.5.

The key advantage for all the methods reviewed in this study is their simple and straightforward ideas. Each method usually combines two or three techniques together to estimate per-pixel classification confidence. The techniques involved in all the methods are well established in the research community. Among all the methods, the geostatistical method and the logistical geographically weighted regression method are a little more
sophisticated than other methods. The geostatistical method and the logistical geographically weighted regression method each produces a probability vector for each class. The probability vector from geostatistical method is further used to construct index for classification confidence whereas that from logistical geographically weighted regression is directly used as classification confidence for each pixel.

Table 2.5 Advantages and limitations of studies on per-pixel classification confidence

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantage</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification score based</td>
<td>• It requires no more effort since classification scores is a byproduct</td>
<td>• The theoretical foundation of the method is not well proved.</td>
</tr>
<tr>
<td>method (CSBM)</td>
<td>of image classification.</td>
<td>• The results of different classifiers may be different and thus not comparable.</td>
</tr>
<tr>
<td>Interpolation based method</td>
<td>• Error matrix is widely used in the remote sensing community.</td>
<td>• It usually requires large number of sample data.</td>
</tr>
<tr>
<td>(IBM)</td>
<td>• Bootstrap is a well established statistical method to estimate</td>
<td>• The assumption of spatially continuous distribution of variables for interpolation may not hold.</td>
</tr>
<tr>
<td></td>
<td>misclassification probability.</td>
<td>• Bootstrap and kriging method may be time consuming.</td>
</tr>
<tr>
<td></td>
<td>• Interpolation techniques are also well accepted by researchers.</td>
<td></td>
</tr>
<tr>
<td>Regression based method</td>
<td>• The idea is straightforward, especially for the global regression</td>
<td>• Data for predictors may not be easily available.</td>
</tr>
<tr>
<td>(RBM)</td>
<td>method.</td>
<td>• Model selection may be difficult.</td>
</tr>
<tr>
<td></td>
<td>• The statistical analysis is routine and mature in statistical fields.</td>
<td></td>
</tr>
</tbody>
</table>
The limitations of the three categories of methods are obvious. Regression based method has two key limitations. First, data for predictors may not be easily available. Second, model selection may be difficult. The interaction of these two limitations usually leads to a serious result: there is either no qualified data for covariates or no significant and meaningful covariates to construct the model.

Interpolation based method also has two key limitations. First, it usually requires a large number of samples well distributed across the whole image. In practice, it would be too expensive or even impossible to satisfy this rigid requirement. Precise interpolation estimates are not guaranteed with sparse training locations (Steele et al. 1998). Second, the assumption of spatially continuous distribution of variables for interpolation may not hold. In practice, the classification confidence may change promptly at the class boundaries (Congalton 1988b).

Compared to regression based method and interpolation based method, classification score based method does not have significant limitations. The advantages of classification score based methods are also promising. First, it requires no more effort since classification scores is a byproduct of image classification. Second, it directly gives the classification confidence at per-pixel.

7.2 Different definitions of per-pixel classification confidence

For convenience, I used the term per-pixel classification confidence all through the above review. Actually, various terms have been used in the literature. Table 2.6 is a summary of different terms along with specific measures to represent per-pixel classification confidence. The first column in Table 2.6 refers to different terms regarded
as alternative definitions to classification confidence. The second column refers to the measures to represent per-pixel classification confidence, i.e., the statistics to calculate per-pixel classification confidence.

Besides classification confidence, the most widely used alternative term is classification uncertainty. Other terms include classification quality, thematic uncertainty, spatial map accuracy, local classification accuracy, spatial classification accuracy, per cell classification accuracy, etc.

Various measures have been used to represent per-pixel classification confidence in the literature. The most widely used measures are posterior probability and relative entropy. Other measures include margin of victory, normalized activation level, average nearest unlike neighbor index, similarity ratio, absolute accuracy, relative accuracy, mixture level, and incompleteness, misclassification probability, uncertainty index, and local classification accuracy, etc.
Table 2.6 Terms and measures for classification confidence

<table>
<thead>
<tr>
<th>Terms</th>
<th>Measures</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification confidence</td>
<td>relative entropy</td>
<td>Maselli et al. (1994)</td>
</tr>
<tr>
<td></td>
<td>relative entropy, posterior probability,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>margin of victory</td>
<td>McIver and Friedl (2001)</td>
</tr>
<tr>
<td></td>
<td>posterior probability</td>
<td>Liu et al. (2004)</td>
</tr>
<tr>
<td></td>
<td>average nearest unlike neighbor index,</td>
<td>Delany et al. (2005)</td>
</tr>
<tr>
<td></td>
<td>similarity ratio, etc.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>distance to second cluster</td>
<td>Mitchell et al. (2008)</td>
</tr>
<tr>
<td>Classification uncertainty</td>
<td>absolute accuracy, relative accuracy,</td>
<td>Shi (1994, 2010)</td>
</tr>
<tr>
<td></td>
<td>mixture level, and incompleteness</td>
<td></td>
</tr>
<tr>
<td></td>
<td>uncertainty factor</td>
<td>Gong et al. (1996)</td>
</tr>
<tr>
<td></td>
<td>posterior probability</td>
<td>Canters (1997)</td>
</tr>
<tr>
<td>Classification uncertainty</td>
<td>Classification uncertainty</td>
<td>Yu et al. (2008)</td>
</tr>
<tr>
<td>Classification quality</td>
<td>posterior probability</td>
<td>Foody et al. (1992)</td>
</tr>
<tr>
<td></td>
<td>normalized activation level</td>
<td>Foody (2000)</td>
</tr>
<tr>
<td>Thematic uncertainty</td>
<td>posterior probability, normalized</td>
<td>Brown et al. (2009)</td>
</tr>
<tr>
<td></td>
<td>activation level</td>
<td></td>
</tr>
<tr>
<td>Spatial map accuracy</td>
<td>misclassification probability</td>
<td>Steele et al. (1998)</td>
</tr>
<tr>
<td>Local/Spatial classification accuracy</td>
<td>uncertainty index</td>
<td>Kyriakidis and Dungan (2001)</td>
</tr>
<tr>
<td>Spatial classification accuracy</td>
<td>local classification accuracy</td>
<td>Foody (2005)</td>
</tr>
<tr>
<td>Per cell classification accuracy</td>
<td></td>
<td>Oort et al. (2004)</td>
</tr>
<tr>
<td>Spatially-explicit classification</td>
<td></td>
<td>Burnicki (2011)</td>
</tr>
<tr>
<td>error probability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spatial variation of classification</td>
<td></td>
<td>Comber et al. (2012)</td>
</tr>
<tr>
<td>accuracy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
7.3 Method validation

Classification confidence at pixel level can be estimated using any method reviewed in this chapter. The validity of the estimated classification confidence should be evaluated with true classification confidence. Previous studies have used different approaches to validate their methods, including visual comparison, simple quantitative comparison, statistical analysis, etc.

7.3.1 Visual comparison

Some studies visually compare the map of classification confidence with classification remote sensing image, class map, or classification accuracy map to verify their results. For example, Gong et al. (1996) classify image with neural network and obtain maximum possibility value similar to the maximum posterior probability from maximum likelihood classification. They then visually compare the uncertainty map with the maximum possibilities of correctly classified grid cells. They find that there is negative relationship between classification uncertainty and maximum possibility value.

In a study using bootstrap method, Steele et al. (1998) compares the map of misclassification probability with remote sensing image and class map. They find that homogeneous areas have low misclassification probability while heterogeneous areas have high misclassification probability.

Similar to Steele et al. (1998), Mitchell et al. (2008) validate their method by visually comparing the confidence map with the remote sensing image and class map. They find that homogeneous areas tend to have high classification confidence while heterogeneous areas with mixed classes tend to have low classification confidence.
7.3.2 Simple quantitative comparison based on binned pixels

In some studies with classification score based method, the estimated classification confidence is validated based on binned pixels. The validation involves those pixels with reference class labels. The detailed validation procedures are explained as follows.

First, grouping pixels. Pixels are grouped into a certain number of bins.

Second, calculating estimated and true mean classification confidence. The estimated mean classification confidence is the average classification confidence of the pixels in each bin. The true mean classification confidence is the proportion of pixels being correctly classified in each bin.

Third, examining relationship between estimated and true mean classification confidence. The two statistics are compared to examine if they are related.

For example, in Maselli et al. (1994), complete coverage reference data are used. All the classified pixels are grouped into five bins according to their relative entropy. The overall accuracy in each bin is compared with the level of relative entropy. It is found that relative entropy is negatively related to classification accuracy.

Similar to Maselli et al. (1994), McIver and Friedl (2001) also group pixels into five bins based on classification confidence. Since they do not have the luxury complete coverage reference data as Maselli et al. (1994), McIver and Friedl (2001) use cross-validation to examine the relationship between classification error and classification confidence. They find that classification confidence measured by probability of the most probable class is positively related to frequency of correctly classified pixels. They also
find that the proportion of misclassified pixels is close to 100% minus mean classification confidence in each bin.

7.3.3 Statistical analysis

Some studies conduct a more statistically rigorous validation of classification confidence other than the simple comparison above. This statistical validation involves the definition of classification confidence. To my best knowledge, there is no clear definition of classification confidence in the literature. Generally, classification confidence is perceived as a statistic measuring the probability of a pixel being correctly classified. Upon accepting this definition, a statistical validation involves calculating measures of agreement and conducting correlation test. Three examples are summarized below.

The first example is referred to Brown et al. (2009), where classification score based method is used. Brown et al. (2009) compare the performance of three classifiers in estimating per-pixel classification confidence. The validation is based on pixels grouped into 10 bins by maximum posterior probability or normalized activation level. The estimated classification confidence is represented with mean probability or activation level while the true classification confidence is represented with the proportion of pixels being correctly classified in each bin. The validation is carried out in two ways: (1) RMSE measuring agreement, and (2) F-test of correlation. They find that all scenarios pass F-test, indicating one-to-one relationship between estimated and true classification confidence. As to the RMSE results, some scenarios have low RMSE while other scenarios have high RMSE. Low RMSE indicates the estimated classification confidence agrees well with the true classification confidence. On the contrary, high RMSE shows
that the estimated classification confidence deviates from the true classification confidence.

The second example is from van Oort et al. (2004), which estimate classification confidence using logistic regression model. The study area is composed of 55 randomly selected subregions. The classification confidence is defined as the probability of correct classification \( p(c) \) for each cell (pixel) and is estimated through the regression model. The true classification confidence is defined as a binary variable \( y(c) \) indicating a cell being correctly or incorrectly classified. van Oort et al. (2004) validate the model through a measure \( SM \). \( SM \) is the absolute difference between \( y(c) \) and \( p(c) \) for all cells,

\[
SM = \sum_{r=1}^{R} \sum_{c=1}^{n_r} |p(c) - y(c)|
\] (2.14)

where \( n_r \) is the number of cells (pixels) in region \( r \), \( R = 55 \) is the total number of regions.

The third example comes from Burnicki (2011), which uses a generalized additive model to estimate change-classification error. Classification confidence is defined as an binary variable indicating change-classification error. The estimated classification confidence is discretized to binary variable through thresholding. The predictive performance of the model is validated using a cross-tabulation matrix of predicted and observed change-classification error.

### 7.3.4 Other approaches

Besides the above three types of validation, other approaches have also been proposed. For example, Foody et al. (1992) find that adding more fieldwork for pixels with low posterior probability can improve classification accuracy.
7.4 Summary

During the past two decades, per-pixel classification confidence has attracted quite a number of studies both from geographical information science and remote sensing. Despite the great progress, several remaining key issues hinder the wide practice of per-pixel classification confidence. As pointed out by Devillers et al. (2010), one metric to assess a research field is its "scientific footprint", i.e., the impact of the field on the non-academic community. A partial failure of spatial data quality research is the poor connection between academic use and day-to-day use of spatial data. This is also the case for per-pixel classification confidence, which has not been widely accepted by users outside the academic community. Besides the limitations listed in Table 2.5, there are three key issues that require further studies.

First, previous studies have not been rigorously validated with empirical evidence. Although some studies have validated their results with certain methods, they are far beyond adequate. Among the four categories of validation approaches reviewed Section 7.3, the statistical validation presented in Section 7.3.3, specifically, the work by Brown et al. (2009), is most promising. Brown et al. (2009) compare the performance of three classifiers, i.e., maximum likelihood classification, and two neural network classifier, PNN and MLP. Brown et al. (2009)'s work can be extended in two points: (1) examining the performance of other commonly used classifiers, including support vector machine and boosted decision tree; (2) validating using complete coverage reference data.

Second, the confusion of concepts in different studies is problematic. There is no commonly accepted terminology in remote sensing image classification and spatial data quality (Caners 1997; Devillers et al. 2010; Jansen and van der Wel 1994). As
demonstrated in Table 2.6, different terms and measures are used in previous studies. Confusion in concepts prevents efficient communication and health development in the field. Therefore, it is imperative to clarify and synthesize concepts.

Third, there is no comprehensive comparison of the main existing methods. Table 2.5 is only a qualitative review of the pros and cons of the existing methods.

In the following two chapters, I will provide a comprehensive evaluation and comparison of the main methods reviewed in this chapter. I will examine the usefulness of different methods and give directions on further studies. Specifically, Chapter 3 evaluates classification score based method; Chapter 4 evaluates interpolation based method. Due to the lack of data, I will not evaluate regression based method in this dissertation. The evaluation of regression based method remains a problem to be solved in future studies.
CHAPTER 3 EVALUATING CLASSIFICATION SCORE BASED METHODS ON PER-PIXEL CLASSIFICATION CONFIDENCE

1 Introduction

Classification score based method derives classification confidence at pixel level using classification scores from classifiers. The most widely used classification scores are the (posterior) probabilities of class membership, also called probability vectors. Besides posterior probabilities, other measures such as activation level of neural network, distance to neighbors or cluster centroids in $k$-NN are also used in the literature. Usually a simple transformation of the classification scores is applied to generate per-pixel classification confidence. The resultant measures include posterior probabilities, weighted posterior probabilities, entropy, normalized maximum possibility, margin of victory, binned average of posterior probabilities and activation level, absolute accuracy, relative accuracy, mixture level, and incompleteness, standardized distance to cluster centroids and related measures, etc..

Classification score based method has attracted continuous attention from scholars both in geographic information sciences and remote sensing. Foody et al. (1992) and Goodchild et al. (1992) are two early studies which use probability vector as an indicator for estimating classification confidence at pixel level. Maselli et al. (1994) proposes a measure named "relative probability entropy" which characterize the heterogeneity in

Besides maximum likelihood classifier, classification score based method has also been applied to non-parametric classifiers such as neural network and boosted decision tree. For example, Gong et al. (1996) propose an uncertainty model using the probability output from neural network classification. Foody (2000) derives entropy using the output activation level also from neural network to measure per-pixel classification confidence. McIver and Friedl (2001) applies classification score based method to image classification with boosted decision tree and finds that classification error is negatively correlated with probability of membership associated with the most probable class.

With the success in applying classification score based method to different classifiers, some researchers have studied the performance of classification score based method in multi-classifier comparison and combination. For example, Liu et al. (2004) proposes a hybrid classifier combining ARTMAP neural network and decision tree to provide per-pixel classification confidence. Brown et al. (2009) compares the pros and cons of maximum likelihood classifier and two types of neural network classifiers in predicting thematic accuracy and classification uncertainty.

Despite the fact that classification score based method has been widely used by researchers, previous studies have not been rigorously validated with empirical evidence. Although some studies have validated their results with certain methods, they are far
beyond adequate. Among the previous studies using classification score based methods, Brown et al. (2009) provide the most comprehensive evaluation. The validation is based on pixels grouped into 10 bins by maximum posterior probability or normalized activation level. The estimated classification confidence is represented with mean probability or activation level while the true classification confidence is represented with the proportion of pixels being correctly classified in each bin. The validation is carried out using two techniques: F-test and RMSE. F-test examine the correlation between estimated classification confidence and true classification confidence while RMSE measures the agreement between the two.

This chapter examines the usefulness of classification scores and derived measures in predicting per-pixel classification confidence. The strength of this chapter lies in two points. First, it examines the performance of four commonly used classifiers: maximum likelihood classifier (MLC), neural network (NN), support vector machine (SVM), boosted decision tree (BDT). Second, it provides a comprehensive evaluation using complete coverage reference data. The rest of the paper is organized as follows. Section 2 is the methodology part. Section 3 is the results, followed by discussion and conclusions in Section 4.

2 Methodology

The methodology of this chapter is illustrated in Figure 3.1. First, remote sensing image is classified to produce class probabilities and class map. Second, class map is compared with reference map to create indicator map for error and correct pixels. Third, probability measures are calculated using class probabilities. Forth, pixels are grouped into bins based on the probability measure, binned classification quality is then calculated.
by averaging pixels in each bin. Fifth, the relationship between binned classification quality and probability measures is examined. The above procedures are applied to three interpolation based methods with four classifiers and three datasets. In total there are $4 \times 3 \times 3 = 36$ scenarios.

2.1 Test datasets

In this chapter, I evaluate classification score based method using three representative datasets with different spatial resolution, spectral bands, and number of classes (Figure 3.2, Table 3.1).
Figure 3.2 Datasets and corresponding reference map

Note: (a) TM image (7-4-2 band composite), (c) QuickBird image (4-3-2 band composite), (e), HYDICE image (63-52-36 band composite); (b), (d), (f) with legend, corresponding reference maps for (a), (c), (e).
Table 3.1 Study site and data

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Study Site</th>
<th>Sensor</th>
<th>Resolution (m)</th>
<th>Acquisition Date</th>
<th>Image size (pixel)</th>
<th>Figure 3.2 subplot Image</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Kent, MD</td>
<td>TM</td>
<td>30</td>
<td>23/1/2010</td>
<td>500×500</td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td>2</td>
<td>Oakland, CA</td>
<td>QuickBird</td>
<td>0.6</td>
<td>24/3/2003</td>
<td>500×500</td>
<td>(c)</td>
<td>(d)</td>
</tr>
<tr>
<td>3</td>
<td>DC Mall</td>
<td>HYDICE</td>
<td>3</td>
<td>23/8/1995</td>
<td>1280×307</td>
<td>(e)</td>
<td>(f)</td>
</tr>
</tbody>
</table>

Complete coverage reference map is developed for each dataset, which is critical for testing and validating classification score based method. The early work of Maselli et al. (1994) to study classification confidence is actually evaluated using complete coverage reference map. Hollister et al. (2004) and Stehman and Wickham (2011) are two recent studies using complete coverage reference map to study classification accuracy but in a non-spatial aspect.

(1) Data 1: Kennedyville cropland, Landsat TM image

Data 1 is a study site in the agricultural heartland (76.00°W, 39.30°N) of Kent County, Maryland. The image spreads across two villages: Worton and Kennedyville. The land is mainly covered with agricultural field. The north side belongs to the Sassafras River watershed while the south side belongs to the Chester River watershed. Forest covers along both rivers. Landsat image (acquisition date: 23 January, 2010) is obtained from USGS Landsat Archives. A clip image of 500×500 pixels is used. Reference map is developed by experienced data analysts and is deemed accurate. Reference map consists of five classes: tree, mature crop, young crop, vacant land, and water.

(2) Data 2: Oakland residential, QuickBird image
Data 2 is about a residential area (122.154° W, 37.780° N) of Oakland, California, where low density residential housing is surrounded by forest. The pan-sharpened multi-spectral QuickBird satellite image is taken with off-nadir viewing angle 11°. A clip image and reference map of 500 × 500 pixels are from Liu and Xia (2010). Reference map is generated by manual interpretation aided with high-resolution (0.3m) USGS orthoimage of the same area and is cross-validated by two interpreters. The reference map consists of six classes: tree, grass, vacant land, road, roof, and shadow.

(3) Data 3: DC Mall, HYDICE image

The image covers the Washington DC Mall and is taken by Hyperspectral Digital Imagery Collection Experiment (HYDICE) sensor (Landgrebe 2003), with the size of 1280 × 307 pixels. Using principle component analysis, I reduce the image dimension to 10 bands which contains 99.9% variance of the original 191 bands. Reference map is generated by manual interpretation aided with high-resolution orthoimage of the same area and is cross-validated by two interpreters. The reference classification map consists of seven classes: tree, grass, trail, road, roof, shadow, and water.

These three datasets will be also be used in Chapter 4-5.

For each dataset, randomly select $S_1 + S_2$ sample pixels, where $S_1$ pixels are used as training set, and the remaining $S_2$ pixels as test set. $S_1$ and $S_2$ is non-overlapped and independent with each other. In this study, I set $S_1 = S_2 = N = 2000$.

2.2 Classifier and classification scores

In the following, I briefly introduce four classifiers used in this study with the focus on deriving classification scores, mainly class probabilities. In previous studies, different researchers have used different classifiers and algorithms. The classification scores
output from classifiers are highly dependent on the classification algorithms used. Therefore, I allocate some space to introduce the development of class probabilities in this section.

The main setting for the classification problem is summarized as follows. Remote sensing image $X$ with $d$ bands is classified into $K$ classes. Each classifier produces a probability vector $\{p_k = P(Y = \omega_k | X)\}$ for each pixel, where $X = (x_1, x_2, \ldots, x_d)$, represents the spectral data of the pixel, $X \in \mathbb{R}^d$, $\omega_k$ is the class label, $k = 1, 2, \ldots, K$, and $K$ is the number of classes. The training data are $Tr = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$, where $X_i \in \mathbb{R}^d$, $Y_i \in \{w_1, \ldots, w_k\}$, $i = 1, 2, \ldots, N$.

The classification is done by assigning each pixel $X$ to the most probable class, i.e.,

$$f(X) = \omega_k \quad \text{if} \quad P(\omega_k | X) > P(\omega_l | X) \quad \text{for all } l \neq k. \quad (3.1)$$

The classification scores used in this study are the class probabilities output from each classifier.

### 2.2.1 Maximum likelihood classifier

Maximum likelihood classifier is the most commonly used supervised classifier in the remote sensing community (Richards and Jia 2006). Maximum likelihood classifier is a type of Bayesian classifier, which directly outputs posterior probability of class membership. Based on Bayes' theorem, the posterior probability of pixel $X$ being class $\omega_k$ is estimated as

$$p_k = P(\omega_k | X) = \frac{P(X | \omega_k)P(\omega_k)}{P(X)} = \frac{P(X | \omega_k)P(\omega_k)}{\sum_{k=1}^{K} P(X | \omega_k)P(\omega_k)}, \quad (3.2)$$
where \( P(\omega_k) \) is the prior probability of class \( \omega_k \) occurs in image \( X \), and

\[
P(X | \omega_k) \quad \text{is the probability distribution or likelihood for class } \omega_k.
\]

In maximum likelihood classification, \( P(X | \omega_k) \) is assumed to follow multivariate normal density function for each class, and is estimated using the training data. In my implementation, the prior probabilities are estimated as the relative frequencies in the training data. According to equation (2.2), the output posterior probabilities sum to one for each pixel.

### 2.2.2 Neural network

Neural networks are nonlinear statistical learning models commonly used in regression and classification (Hastie et al. 2009). The most widely used neural network is the single hidden layer back-propagation network, which is composed of three layers: input layer, hidden layer, and output layer (Figure 3.3). Each layer is composed of certain number of basic units called nodes. The number of input nodes is usually determined by the number of features (bands) of the remote sensing image while the number of output nodes is generally the same as the number of training classes. The number of hidden nodes is determined based on background knowledge and experimentation, which could start with the number of input nodes to some large numbers (Hastie et al. 2009; Richards and Jia 2006). Besides the commonly used single hidden layer algorithm, multiple hidden layers can also be designed to construct hierarchical features.
The input information $X$ is processed through activation function $\sigma(\nu)$, and is then combined to derive target $Y$. The most commonly used activation function is the sigmoid activation function, $\sigma(\nu) = 1 / (1 + e^{-\nu})$. Another activation function, though not as widely used as the sigmoid function, is radial basis function (RBF). Based on the structure in Figure 3.3, we have

$$
Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), \quad m = 1, \ldots, M,
$$

$$
T_k = \beta_{0k} + \beta_k^T Z, \quad k = 1, \ldots, K,
$$

$$
Y = f_k(x) = g_k(T), \quad k = 1, \ldots, K,
$$

where $X = (X_1, X_2, \ldots, X_d)$, $Z = (Z_1, Z_2, \ldots, Z_M)$, $T = (T_1, T_2, \ldots, T_K)$. Now we have

$$
\{\alpha_{0m}, \alpha_m; m = 1, 2, \ldots, M\} \; M(d + 1) \; \text{weights},
$$

$$
\{\beta_{0k}, \beta_k; k = 1, 2, \ldots, K\} \; K(M + 1) \; \text{weights}.
$$

These weights are estimated by fitting the training data using back-propagation approach.

I implement Multi-layer Perceptron (MLP) with sigmoid activation function in this paper. Multi-layer perceptron is the most widely used neural network in remote sensing (Atkinson and Tatnall 1997; Brown et al. 2009). The output activation level of multi-layer perceptron has been used to generate per-pixel classification confidence in previous
studies (Brown et al. 2009; Foody 2000; Gong et al. 1996). The network is trained with scaled conjugate gradient algorithm (Moller 1993), which provides faster convergence than standard back-propagation algorithm (Hastie et al. 2009). Main factors influencing the performance of multi-layer perceptron include network architecture, training dataset, and number of iteration. In my experiment, the number of input nodes equals the number of image bands $d$, and the number of output nodes is set to the number of class. I have one hidden layer, and the number of hidden nodes is 20 to accommodate test accuracy and computation time. The largest number of iteration is set to 1000. Class probabilities are calculated by normalizing the output activation level to sum to one.

2.2.3 Support vector machine

Support vector machine is a kernel-based learning method commonly used in machine learning (Burges 1998; Cristianini and Shawe-Taylor 2000; Vapnik 1998). Support vector machine has been successfully applied to classification problems in remote sensing and is found to be competitive with the best available machine learning algorithms (Huang et al. 2002; Melgani and Bruzzone 2004). Support vector machine is originally developed as a binary classifier. There are a bunch of strategies to convert a binary classification to multiclass classification using support vector machine, among which one-against-all and one-against-one are the two most widely used approaches (Hsu and Lin 2002; Liu et al. 2006).

The standard support vector machine does not provide class probabilities (Platt 2000). A number of methods have been developed to generate probability output for support vector machine. For binary support vector machine output, Platt (2000) proposes an approach by fitting a sigmoid function to the discriminant function. Later, Lin et al.
(2007) proposes an improved algorithm which converges theoretically and avoids numerical difficulties of Platt (2000). For multiclass support vector machine, Duan et al. (2003) propose a method by combing binary classifiers using soft-max functions. Specifically for one-against-one support vector machine, several authors (Friedman 1996; Hastie and Tibshirani 1998; Knerr et al. 1990; Price et al. 1995; Refregier and Vallet 1991; Wu et al. 2004) have developed methods to combine pairwise class probabilities. For a detailed review, please refer to Wu et al. (2004).

I implement support vector machine using LIBSVM (Chang and Lin 2011), one of the most widely used support vector machine software. LIBSVM implements one-against-one approach for multiclass classification, which is a competitive approach compared with other methods (Hsu and Lin 2002). As to probability estimates, LIBSVM applies the second approach proposed in Wu et al. (2004), which is more stable than other existing methods. There are two steps to estimate the probability output. First, estimate the pairwise class probabilities

\[ r_{kl} = P(y = \omega_k \mid y = \omega_k \text{ or } \omega_l, X), \quad (3.5) \]

Following Lin et al. (2007), let \( \hat{f} \) be the decision value at \( x \), and assuming

\[ r_{kl} \approx P(y = \omega_k \mid y = \omega_k \text{ or } \omega_l, X) = \frac{1}{1 + e^{A_f+B}}, \quad (3.6) \]

\( r_{kl} \) can be estimated by minimizing the negative log likelihood of training data.

Second, \( p_k \) is derived by combining \( r_{kl} \) through solving a coupled pairwise optimization system
$$\min_p \frac{1}{2} \sum_{k=1}^{K} \sum_{l \neq k} (r_{lk} p_k - r_{kl} p_l)^2,$$

s.t. \( p_k \geq 0, \forall k, \sum_{k=1}^{K} p_k = 1. \) \hspace{1cm} (3.7)

I use radial basis function kernel in my classification, and select penalty parameter \( C \) and Gaussian parameter \( \gamma \) using grid search algorithm and five-fold cross validation. For details of implementing LIBSVM, please refer to Chang and Lin (2011). As indicated in Equation (2.5), the output probabilities of class membership are normalized so that they sum to one for each pixel.

### 2.2.4 Boosted decision tree

Boosting is a powerful learning technique which generates strong ensemble learner through combining a set of weak learners (Freund and Schapire 1997). Boosting mostly use decision tree, e.g., CART, C4.5, or C5.0, as the weak learner although it has also been applied to other classifiers including neural network (Schwenk and Bengio 2000). The AdaBoost.M1 proposed by Freund and Schapire (1997) is the most widely used discrete boosting algorithm, in which the base classifier returns a discrete class label. Friedman et al. (2000) proposes a generalized version called Real AdaBoost which provides class probability estimates. Collins et al. (2002) and Schapire et al. (2005) propose a logistic version of AdaBoost, called AdaBoost.L, which unifies boosting with logistic regression based on the work of Friedman et al. (2000) and others. Although boosting has been successfully applied to remote sensing image classification (e.g., Briem et al. 2002; McIver and Friedl 2001; Pal and Mather 2003; Waske and Braun 2009), it is not well known to the remote sensing community. Therefore, I present the boosting algorithm with more detail than other classifiers.
1. Input: \((X_1, Y_1), \ldots, (X_N, Y_N); \ X_i \in \mathcal{X} \ \omega_1, \ldots, \omega_k \} \)

2. Initialize: \(W_1(i) = 1/N. \)

3. For \(k = 1, \ldots, K \):
   \[
   Y'_i = \begin{cases} 
   +1, & \text{if } Y'_i = \omega_k, \\
   -1, & \text{otherwise.}
   \end{cases}
   \]
   (One-against-All)

   Let \(\lambda_i(i) = 0. \)

For \(t = 1, \ldots, T: \)

- Train weak learner using distribution \(W_t. \)
- Get weak classifier \(G_t. \)

   The class probability at each tree node \(s: \)
   \[
   p^+_s = P(+1|s), \quad p^-_s = P(-1|s) .
   \]

   Compute: \(\delta_{t,s} = \frac{1}{2} \ln \left( \frac{p^+_s}{p^-_s} \right), \quad \delta_t(i) = \delta_{t,s(\{x,y\})} \)

   - Update: \(\lambda_{t+1}(i) = \lambda_t(i) + \delta_t(i). \)
   - Update: \(W_{t+1}(i) = \frac{1}{1 + \exp(Y'_i \lambda_{t+1}(i))}, \) and normalize \(W_{t+1}. \)

4. Output the final classifier: \(G(X) = \sum_{t=1}^{T} G_t(X). \)

5. Get The posterior probability: \(\hat{P}(\omega_k | X) = \frac{1}{1 + \exp(-G(X))}. \)

---

Figure 3.4 AdaBoost.ML: a multiclass logistic version of AdaBoost

Note: Designed based on ideas from Collins et al. (2002) and Schapire et al. (2005).

In this paper, I implement a multiclass logistic version of AdaBoost with CART as the weak learner (Figure 3.4). The binary classification algorithm AdaBoost.L is generalized to multiclass using one-against-all procedure. In the training process, CART
is sequentially applied to repeatedly modified version of the training data, thereby producing a sequence of weak classifier $G_t(X)$, $t = 1, 2, \ldots, T$. In each iteration $t$, training samples are modified, giving higher weight to cases that are currently misclassified. The final classifier is a linear combination of $G(X) = \sum_{t=1}^{T} G_t(X)$. The posterior probabilities are estimated as $P(\omega_k \mid X) = \frac{1}{1 + \exp(-G(X))}$, which are then normalized to sum to one.

The procedure to weight samples in the boosting process (Step 3 in Figure 3.4) requires iterative modification of the weak classifier. Some researchers avoid modifying weak learner by resampling proportional to weights (McIver and Friedl 2001). However, two limitations exist in this resampling practice: (1) modest rounding error is introduced (McIver and Friedl 2001); (2) randomization in resampling induce uncertainty to the results (Friedman et al. 2000). In order to overcome these two issues, I follow the original boosting to weight training samples rather than resampling.

Main parameters affecting the performance of boosted decision tree include the number of nodes and the number of iterations. The number of nodes determines the complexity of the classification tree and is related to the strength of the base learner. The base classifier is very weak when there are only a few nodes. However, the base learner may overfit the data if there are too many nodes. The number of iteration directly influence the computation time. I apply a grid-search algorithm to find the optimal combination of these two factors. Generally, the number of iteration is around 30, while the maximum number of nodes is around 32.
2.3 Probability measures derived from classification scores

Three measures are derived using the classification scores from classifiers. The classification scores are obtained from four classifiers using the algorithms in Section 2.2. These classification scores take the form of probability values ranging between 0-1. Therefore, I name the measures derived from classification scores as probability measures. Recall that the image is classified into $K$ classes and the classifier outputs class probabilities for each pixel. Without loss of generality, rank the non-zero class probabilities for a pixel as

$$p_1 \geq p_2 \geq \ldots p_{k} \geq \ldots \geq p_{K'} > 0,$$

where

$$p_k = P(\omega_k | X), \quad \sum_{k=1}^{K'} P(\omega_k | X) = 1, \quad k = 1, 2, \ldots, K' \leq K.$$

The following three measures are generally used in the literature:

(1) Primate probability (PP): the probability of membership associated with the most probable class: $p_1$. In image classification, each pixel is usually assigned the class label of primate probability (Richards and Jia 2006). While the term "primate probability" is coined by the author, other names have been defined in the literature. For example, McIver and Friedl (2001) define it as "classification confidence". Primate probability ranges between $1/K$ and 1, and its extrema are determined as follows:

- $\text{PP}_{\text{min}} = 1/K$, when $p_1 = p_2 = \ldots = p_K = 1/K$.
- $\text{PP}_{\text{max}} = 1$, when $p_1 = 1$ and $p_k = 0$ for all $1 < k \leq K'$.

(2) Margin of victory (MV): According to McIver and Friedl (2001), margin of victory is the difference in the class probabilities between the most probable class and the second most probable class:
Margin of victory ranges between 0 and 1, and its extrema are determined as follows:

- \( \text{MV}_{\min} = 0, \) when \( p_1 = p_2. \)
- \( \text{MV}_{\max} = 1, \) when \( p_1 = 1 \) and \( p_k = 0 \) for all \( 1 < k \leq K'. \)

(3) Relative entropy (RH): Relative entropy is first proposed by Maselli et al. (1994) and is widely used in the literature. Relative entropy is derived from information entropy:

\[
\text{RH} = \frac{H}{H_{\max}} = \frac{-\sum_{k=1}^{K'} p_k \log p_k}{\log K},
\]

where \( H \) is the information entropy of a pixel. Relative entropy ranges between 0 and 1, and its extrema are determined as follows:

- \( \text{RH}_{\min} = 0, \) when \( p_1 = 1 \) and \( p_k = 0 \) for all \( 1 < k \leq K', \)
- \( \text{RH}_{\max} = 1, \) when \( p_1 = p_2 = \ldots = p_K = 1/K. \)

2.4 Approaches to evaluate method performance

2.4.1 Bi-histogram

The bi-histogram is a exploratory analysis tool for examining the distribution of two datasets by the histograms of both datasets. The bi-histogram is powerful in that it shows the distribution features such as location, scale, skewness, and outliers of both datasets in a single plot. In this study, I use bi-histogram to explore the relationships between probability measures for correct and error pixels.
2.4.2 Binned classification quality

Previous studies propose a method to develop a statistic by grouping pixels into bins. I name this statistic as binned classification quality (BCQ). The procedures to calculate binned classification quality are the same for primate probability, margin of victory, and relative entropy. Therefore, the following shows the steps to calculate binned classification quality based on primate probability for demonstration purpose.

First, divide pixels of the image into a set of $Q$ bins based on primate probability. Let $M_{\text{min}}$ and $M_{\text{max}}$ be the minimum and maximum values for primate probability, the range $[M_{\text{min}}, M_{\text{max}}]$ is divided into $Q$ intervals:

$$M_q = \begin{cases} 
(b_{q-1}, b_q], & \text{when } q = 2, 3, \ldots, Q, \\
[b_0, b_1], & \text{when } q = 1.
\end{cases}$$

(3.11)

Here I have $b_0 = M_{\text{min}}, b_Q = M_{\text{max}}$. Break points $b_q$ are defined by dividing the interval of $[M_{\text{min}}, M_{\text{max}}]$ with equal distance, i.e.,

$$b_q = M_{\text{min}} + q\Delta,$$

(3.12)

where $\Delta = \frac{M_{\text{max}} - M_{\text{min}}}{Q}$ is the length of the bin, $q = 1, 2, \ldots, Q$.

Second, binned classification quality is estimated as the proportion of correct pixels in each bin. Binned classification quality is actually the overall accuracy based on pixels belonging to each bin. The difference of the lower ($b_{q-1}$) and upper ($b_q$) bound, i.e., $b_q - b_{q-1}$, for each bin is very small. It is assumed that pixels in each bin $M_q$ have the same primate probability equal to $b_q$, and same per-pixel classification confidence equal to BCQ$_q$. Using the procedures above, two sequences of data, BCQ$_q$ and $b_q$, are obtained for primate probability.
Based on previous studies (Brown et al. 2009; Maselli et al. 1994; McIver and Friedl 2001), the hypothesis to test is: Classification scores can predicate classification quality. Specifically, the hypothesis is explained as: Correct pixels have high primate probability, high margin of victory, and low relative entropy. Therefore, I examine the following hypothesis:

- For primate probability, \( BCQ_q \) has positive relationship with \( b_q \).
- For margin of victory, \( BCQ_q \) has positive relationship with \( b_q \).
- For relative entropy, \( BCQ_q \) has negative relationship with \( b_q \).

I conduct correlation analysis between \( BCQ_q \) vs. \( b_q \) to evaluate the above hypothesis.

### 2.5 Summary of methodology

I first conduct exploratory data analysis (EDA) using bi-histogram of probability measure for correct and error pixels. I then examine the relationship between binned classification quality and probability measures. The detailed steps are explained as follows.

Step 1: Image classification. The three remote sensing images are classified using four commonly used classifiers: maximum likelihood classifier, neural network, support vector machine, and boosted decision tree. There are in total \( 3 \times 4 = 12 \) scenarios of image classification.

Step 2: Creation of classification confidence map. There are two types of classification confidence map. One is the map of probability measures where each pixel takes the form of a continuous value. The other is the indicator map of classification error/correct where each pixel \( x \) has an indicator value \( I(x) \). \( I(x) = 0 \) when the pixel \( X \) is
correctly classified. \( I(X) = 1 \) when it is incorrectly classified. The indicator map is created by comparing the class map with the full coverage reference map.

Step 3: Calculation of the statistic binned classification quality. Binned classification quality is calculated by binning pixels based on each of the three probability measures, i.e., primate probability, margin of victory, and relative entropy.

Step 4: Method evaluation. I evaluate the usefulness of three probability measures in predicking classification quality using exploratory data analysis tools (q-q plot) and correlation analysis of \( BCQ_q \) vs. \( b_q \).

The results from three probability measures are also compared to examine their difference in predicting the spatial variation of classification quality.

3 Results

3.1 Primate probability

3.1.1 Maps of classification confidence

Figure 3.5 - Figure 3.12 show the spatial distribution of primate probability and classification error for three datasets. The findings are summarized as follows.
Figure 3.5 Maps of primate probability based on MLC

Figure 3.6 Maps of primate probability based on NN
Figure 3.7 Maps of primate probability based on SVM

Figure 3.8 Maps of primate probability based on BDT
Figure 3.9 Maps of classification error based on MLC

Figure 3.10 Maps of classification error based on NN
Figure 3.11 Maps of classification error based on SVM

Figure 3.12 Maps of classification error based on BDT
(1) While the general patterns look similar for different classifiers, there are some minor differences across classifiers.

(2) The spatial patterns of primate probability and classification error match well.

(3) Characteristics for the spatial patterns: For Data 1, there are some patch clusters of low primate probability; For Data 2 and 3, pixels with low primate probability mainly concentrate in roof and road area which reveals the confusion in these two classes.

3.1.2 Comparing primate probability for correct and error pixels

Figure 3.13 shows the bi-histogram of primate probability for correct and error pixels. Generally, most correct pixels have high primate probability values (>0.8), while error pixels have primate probability spread across the whole range of \([1/K,1]\). Here \(K\) is the number of classes.

The distribution of primate probability for correct and error pixels varies with datasets and classifiers.

(1) Primate probability for correct pixels

For correct pixels, the distribution patterns of primate probability can be divided into two groups. First, for Data 1, most correct pixels have primate probability greater than 0.9. Second, for Data 2 and 3, while the majority of the correct pixels have primate probability greater than 0.9, there are more pixels spread across the whole range of \([1/K,1]\) than that of Data 1.

(2) Primate probability for error pixels

For error pixels, the distribution patterns of primate probability can also be divided into two groups. First, all the datasets except Data 3 with support vector machine: primate
probability has similar quasi-uniform distribution pattern. Second, Data 3 with support vector machine: primate probability is approximately a normal distribution.

In summary, if the primate probability of a pixel is high, it is mostly likely a correct pixel. Otherwise, if the primate probability of a pixel is median or low, it is not sure whether it is a correct or error pixel.

Figure 3.13 Bi-histogram of primate probability (PP) for correct and error pixels

Note: y - PP; x: proportion of pixels. Left side - correct pixel, right side - error pixel.
3.1.3 Binned classification quality vs. primate probability

Figure 3.14 shows the scatter plots of BCQ\_q vs. b\_q for primate probability based on all the pixels in each image. There is positive relationship between BCQ\_q vs. b\_q. The exact relationship varies with datasets and classifiers. There are also some outliers when primate probability is small.

Figure 3.15 shows the scatter plots of BCQ\_q vs. b\_q for primate probability based on S\_2 = 2000 test pixels in each image. The positive relationship still exists, but the scatter plots are much more scattered than those in Figure 3.14. For some scenarios, e.g., Data 1 with neural network and boosted decision tree, there are more outliers.

![Figure 3.14 BCQ\_q vs. b\_q based on primate probability for all the pixels of the whole map](image)
Figure 3.15 BCQ_q vs. b_q based on primate probability for 2000 test pixels (S_2)

Table 3.2 shows the correlation coefficient (R) of BCQ_q vs. b_q for primate probability. When the whole map is considered, most cases except Data 3 with neural network, have R greater than 0.95. As to the S_2 test pixels, the correlation coefficient R is smaller than those from the whole map. They are mostly above 0.85 except Data 1 with neural network, support vector machine, and boosted decision tree.
Table 3.2 Correlation coefficients (R) of BCQ\(_q\) vs. \(b_q\) for primate probability

<table>
<thead>
<tr>
<th></th>
<th>Whole map</th>
<th></th>
<th></th>
<th></th>
<th>(S_2) test pixels</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLC</td>
<td>NN</td>
<td>SVM</td>
<td>BDT</td>
<td>MLC</td>
<td>NN</td>
<td>SVM</td>
<td>BDT</td>
</tr>
<tr>
<td>Data 1</td>
<td>0.9953</td>
<td>0.9867</td>
<td>0.986</td>
<td>0.9935</td>
<td>0.8995</td>
<td><strong>0.6825</strong></td>
<td><strong>0.7872</strong></td>
<td><strong>0.8498</strong></td>
</tr>
<tr>
<td>Data 2</td>
<td>0.9917</td>
<td>0.9915</td>
<td>0.9688</td>
<td>0.9779</td>
<td>0.9153</td>
<td>0.9230</td>
<td><strong>0.8419</strong></td>
<td>0.8857</td>
</tr>
<tr>
<td>Data 3</td>
<td>0.9855</td>
<td><strong>0.9075</strong></td>
<td>0.9935</td>
<td>0.9967</td>
<td>0.9134</td>
<td>0.9206</td>
<td>0.9002</td>
<td>0.9138</td>
</tr>
</tbody>
</table>

3.2 Margin of victory

3.2.1 Maps of classification confidence

Figure 3.16 - Figure 3.19 show the spatial distribution of margin of victory for three datasets based on four classifiers respectively. There exist spatial patterns for all three datasets, which is similar to that of primate probability. Compared to Figure 3.5 - Figure 3.8, the cluster of low values in Figure 3.16 - Figure 3.19 are more evident and consistent with the classification errors in Figure 3.9- Figure 3.12.
Figure 3.16 Maps of margin of victory based on MLC

Figure 3.17 Maps of margin of victory based on NN
Figure 3.18 Maps of margin of victory based on SVM

Figure 3.19 Maps of margin of victory based on BDT
### 3.2.2 Comparing margin of victory for correct and error pixels

Figure 3.20 shows the bi-histogram for margin of victory. The distribution patterns are similar to that of the primate probability. Generally, most correct pixels have high margin of victory values above 0.8, while error pixels have margin of victory spread across the whole range of [0,1].

![Bi-histogram of margin of victory (MV) for correct and error pixels](image)

Note: y - MV; x: proportion of pixels. Left side - correct pixel, right side - error pixel.
The distribution of margin of victory for correct and error pixels varies with datasets and classifiers.

(1) Margin of victory for correct pixels

For correct pixels, the distribution patterns of margin of entropy can be divided into three groups. First, for Data 1 with all four classifiers, Data 2 with boosted decision tree, Data 3 with maximum likelihood classifier and boosted decision tree, margin of victory are mostly greater than 0.9. Second, for Data 2 with support vector machine, margin of victory are mostly greater than 0.8. Third, for Data 2 with maximum likelihood classifier, neural network, Data 3 with neural network, support vector machine, margin of victory is negatively skewed but more spread across the whole range of [0,1].

(2) Margin of victory for error pixels

Margin of victory for error pixels has similar distribution patterns with primate probability, which can be divided into two groups. First, for all three datasets except Data 3 with support vector machine, margin of victory has similar quasi-uniform distribution patterns. Second, for Data 3 with support vector machine, primate probability is positively skewed which indicating most pixels having low margin of victory.

In summary, if a pixel has high margin of victory, it is high probable that it is a correct pixels. On the other hand, if a pixel has low margin of victory, it is also high probable that it is an error pixel. However, if a pixel has median margin of victory, we are not sure whether it is a correct or error pixel.

3.2.3 Binned classification quality vs. margin of victory

Figure 3.21 shows the scatter plots of BCQ_q vs. b_q for margin of victory where margin of victory is divided into Q = 30 bins with equal distance. There is positive
relationship between $BCQ_q$ vs. $b_q$. Similar to that of primate probability, the exact relationship varies with datasets and classifiers.

Figure 3.22 shows the scatter plots of $BCQ_q$ vs. $b_q$ for margin of victory based on $S_2 = 2000$ test pixels in each image. Again, the positive relationship still exists, but the scatter plots are much more scattered than those in Figure 3.21. For some scenarios, e.g., Data 1 with sub vector machine, there are more outliers.

Figure 3.21 $BCQ_q$ vs. $b_q$ based on margin of victory for all the pixels of the whole map
Figure 3.22 BCQ\textsubscript{q} vs. \(b_q\) based on margin of victory for 2000 test pixels (\(S_2\))

Table 3.3 Correlation coefficients (R) of BCQ\textsubscript{q} vs. \(b_q\) for margin of victory

<table>
<thead>
<tr>
<th></th>
<th>Whole map</th>
<th>(S_2) test pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLC</td>
<td>NN</td>
</tr>
<tr>
<td><strong>Data 1</strong></td>
<td>0.9991</td>
<td>0.9980</td>
</tr>
<tr>
<td><strong>Data 2</strong></td>
<td>0.9887</td>
<td>0.9926</td>
</tr>
<tr>
<td><strong>Data 3</strong></td>
<td>0.9913</td>
<td>0.9986</td>
</tr>
</tbody>
</table>

Table 3.3 shows the correlation coefficient (R) of BCQ\textsubscript{q} vs. \(b_q\) for margin of victory. When the whole map is considered, all the cases have correlation coefficient
R>0.97, indicating strong positive correlation between BCQ$_q$ vs. $b_q$. As to the $S_2$ test pixels, the correlation coefficient R is smaller than those from the whole map. They are mostly above 0.85 except Data 1 with neural network.

Compared to Table 3.2, the correlation between BCQ$_q$ and $b_q$ for margin of victory is higher than that for primate probability.

3.3 Relative entropy

3.3.1 Maps of classification confidence

Figure 3.23 - Figure 3.26 show the spatial distribution of relative entropy for three datasets based on four classifiers respectively. There exist spatial patterns which look similar with those for primate probability and margin of victory. The clusters of high values in Figure 3.23 - Figure 3.26 generally match with the classification errors in Figure 3.9 - Figure 3.12.
Figure 3.23 Maps of relative entropy based on MLC

Figure 3.24 Maps of relative entropy based on NN
Figure 3.25 Maps of relative entropy based on SVM

Figure 3.26 Maps of relative entropy based on BDT
3.3.2 Comparing relative entropy for correct and error pixels

Figure 3.27 shows the bi-histogram of relative entropy for correct and error pixels. The distribution patterns are quite different from those for primate probability and margin of victory.

Figure 3.27 Bi-histogram of relative entropy (RH) for correct and error pixels

Note: y - RH; x: proportion of pixels. Left side - correct pixel, right side - error pixel.

(1) Relative entropy for correct pixels

For correct pixels, the distribution patterns of relative entropy can be divided into two groups. For Data 1 with all classifiers, relative entropy mostly concentrates at low
values. For Data 2 and Data 3, relative entropy is positively skewed and is more spread out than Data 1.

(2) Relative entropy for error pixels

In most cases, relative entropy approximates the normal distribution with mean values around 0.4-0.5.

In summary, if a pixel has low relative entropy, the chance of it being a correct pixel is high. However, if a pixel has median or high relative entropy, the chance of it being a correct pixel is not sure.

3.3.3 Binned classification quality vs. relative entropy

Figure 3.28 shows the scatter plots of BCQ\textsubscript{q} vs. b\textsubscript{q} for relative entropy where relative entropy is divided into \( Q = 30 \) bins with equal distance. There is generally negative relationship between BCQ\textsubscript{q} vs. b\textsubscript{q}. However, the relationship is not monotonic for Data 2 with maximum likelihood classifier, neural network, and support vector machine. In other words, when relative entropy is very small or very high, there are some outliers in the scatter plots.

Figure 3.29 shows the scatter plots of BCQ\textsubscript{q} vs. b\textsubscript{q} for relative entropy based on \( S_2 = 2000 \) test pixels in each image. Again, the negative relationship still exists, but the scatter plots are much more scattered than those in Figure 3.28. For all three datasets with neural network, Data 1 and Data 2 with support vector machine, there are some outliers in the scatter plots.

Table 3.4 shows the correlation coefficient (R) of BCQ\textsubscript{q} vs. b\textsubscript{q} for relative entropy. When the whole map is considered, all the cases except Data 2 with neural network, support vector machine, and boosted decision tree, have correlation coefficient R < -0.95,
indicating strong negative correlation between $\text{BCQ}_q$ vs. $b_q$. As to the $S_2$ test pixels, the absolute value of correlation coefficient R is smaller than those from the whole map. The correlation coefficients R are mostly smaller than -0.85 except for all three datasets with neural network, and Data 1 and Data 2 with support vector machine.

Figure 3.28 $\text{BCQ}_q$ vs. $b_q$ based on relative entropy for all the pixels of the whole map
Figure 3.29 BCQ$_q$ vs. $b_q$ based on relative entropy for for 2000 test pixels ($S_2$)

Table 3.4 Correlation coefficients (R) of BCQ$_q$ vs. $b_q$ for relative entropy

<table>
<thead>
<tr>
<th>Data</th>
<th>Whole map</th>
<th>$S_2$ test pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLC</td>
<td>NN</td>
</tr>
<tr>
<td>Data 1</td>
<td>-0.9914</td>
<td>-0.9816</td>
</tr>
<tr>
<td>Data 2</td>
<td>-0.9810</td>
<td>-0.9463</td>
</tr>
<tr>
<td>Data 3</td>
<td>-0.9871</td>
<td>-0.9758</td>
</tr>
</tbody>
</table>
3.4 Comparison of three probability measures

In Sections 3.1-3.3, I have examined relationship between three probability measures and binned classification quality. Primate probability and margin of victory are both positively related to binned classification quality, while relative entropy is negatively related to binned classification quality. The correlation for margin of victory is the highest among the three probability measures while relative entropy has the lowest correlation with binned classification quality. There are more outliers on the scatter plots based on primate probability and relative entropy. Especially the scatter plots based on relative entropy, the relationship is not monotonic.

In general, margin of victory is most promising in characterizing per-pixel classification confidence.

4 Discussion and Conclusions

In this chapter, I examined the usefulness of using classification scores to predict classification confidence at pixel level. I experimented three commonly used probability measures derived from the probability output of four classifiers.

First, the spatial patterns of three probability measures are visually similar. The cluster patches of low primate probability, low margin of victory, and high relative entropy matches the maps of classification error.

Second, the bi-histogram of three probability measures shows different distribution patterns. For primate probability and margin of victory, correct pixels and error pixels are distinguishable: the primate probability and margin of victory for correct pixels usually concentrated in high values while that for error pixels is much spread across the whole
value range. The relative entropy is less distinguishable for correct and error pixels in the bi-histogram.

Third, the scatter plots of binned classification quality vs. probability measures show that binned classification quality is generally positively related to primate probability, margin of victory, and negatively related to relative entropy. Margin of victory is more stable in the relationship with binned classification quality while relative entropy is least stable. Therefore, margin of victory seems to be the most promising measure for identifying per-pixel classification confidence.

Forth, it should be pointed out that, the relationship between binned classification quality and probability measures are examined using the full coverage reference data. In practice, we usually do not have this luxury full coverage reference data. In such case, we can only calculate binned classification quality based on test pixels. The number of pixels used to calculate binned classification quality will be much smaller than those used in this chapter. Therefore, the relationship between binned classification quality and probability measure may be more variable in practice.
CHAPTER 4 EVALUATING THREE INTERPOLATION BASED METHODS ON PER-PIXEL CLASSIFICATION CONFIDENCE

1 Introduction

Interpolation based method generates per-pixel classification confidence by interpolating estimation at sample locations to the whole image. In the literature, there are three different approaches which I name as: local error matrix method (Foody 2005), bootstrap method (Steele et al. 1998), and geostatistical method (Kyriakidis and Dungan 2001). In Foody (2005), local classification accuracies are derived based on local error matrices constrained in a local neighborhood. The local accuracies are then interpolated to the whole map using inverse distance squared weighted interpolation, i.e., IDW. In Steele et al. (1998), misclassification rates at sample pixels are obtained through bootstrap resampling. The misclassification rates are then interpolated to the whole map using kriging. Kyriakidis and Dungan (2001) use a more sophisticated method to derive local confusion index from global error matrix. The confusion index is then interpolated to the whole map using simple kriging. The above three methods provide insight on alternative approaches to estimate per-pixel classification confidence.

Despite the fact that these three interpolation based methods have been recognized and cited by various studies (see e.g., Burnicki 2011; Comber et al. 2012; Foody 2002; Stehman and Czaplewski 2003; van Oort 2007), they have not been rigorously evaluated
and tested. In other words, the estimated per-pixel classification confidence has not been evaluated with the true per-pixel classification confidence. An untested method has no warrantee to be correct and effective in practice.

In this chapter, I evaluate the effectiveness of the above three methods using carefully selected representative datasets. The organization of this chapter is summarized as follows. Section 2 describes the methodology for this study which is further divided into four subsections. Section 3 presents the results. Section 4 discusses the results and concludes this chapter.

2 Methodology

The methodology of this chapter is illustrated in Figure 4.1. First, remote sensing image is classified to produce class map and estimate classification confidence map (more detail to follow). Second, class map is compared with reference map to create indicator map for error and correct pixels. Third, pixels are grouped into bins based on classification confidence, and binned classification quality is then calculated by averaging pixels in each bin. Forth, the relationship between binned classification quality and classification confidence is examined. The above procedures are applied to three interpolation based methods with four classifiers and three datasets. In total there are $3 \times 4 \times 3$ scenarios.

The detailed procedures to estimate classification confidence map will be explained in Section 2.2.1-2.2.3 under each of the three interpolation based methods.
2.1 Test datasets

The test datasets used in this chapter are the same as those introduced in Section 2.1 of Chapter 3. Please refer to it accordingly.

2.2 Three interpolation based methods

In this section, I briefly introduce three interpolation based methods to estimate per-pixel classification confidence. All three methods use the same training and test samples. Randomly select $S_1 + S_2$ sample pixels, where $S_1$ pixels are used as training set, and the remaining $S_2$ pixels as test set. $S_1$ and $S_2$ is non-overlapped and independent with each other. In this study, I set $S_1 = S_2 = N = 2000$. 
2.2.1 Local error matrix method

Foody (2005) proposes a method to estimate local classification accuracy by combining local error matrix and interpolation. For convenience, I name this method as local error matrix method (LEM). By partitioning test samples based on sub-regions of the area, local classification accuracy is estimated using test data in each sub-region. The local overall accuracies are then interpolated to the whole map and we obtain local classification accuracy (LCA) at pixel level. Suppose we have a rectangular study area composed of pixels with \( a \) rows and \( b \) columns, i.e., \( a \times b \) pixels in total. The process to estimate local classification accuracy involves the following steps.

Step 1: Image classification.

Based on \( S_1 \) training pixels, the remote sensing image is classified in the usual way to produce a class map. Global error matrix and related accuracy indices, overall accuracy, producer's accuracy, user's accuracy, and kappa, are derived based on the class map and test data. The class map will be further used to create indicator classification confidence map in Section 2.3.1.

Step 2: Definition of grid pixel.

A tessellation of \( m \times n \) grid is created and overlaid with the class map. Name the pixel at the location of grid intersection as grid pixel (See Figure 4.2 for an example with Data 3). In total there are \( m \times n \) grid pixels.

Step 3: Selection of neighbor pixels for each grid pixel.

For each grid pixel, a set of \( k \) neighbor pixels are selected from the test dataset. The neighbor pixels can either be selected using \( k \)-nearest-neighbor (\( k \)-NN) method, or
selected from a fixed size window, e.g., fixed radius method. Figure 4.3 shows five example grid pixels and their $k$ nearest neighbors.

![Grid pixels overlaying with reference map](image1)

Figure 4.2 Grid pixels overlaying with reference map

![Example grid pixels and $k$ nearest neighbors](image2)

Figure 4.3 Example grid pixels and $k$ nearest neighbors

Step 4: Calculating local classification accuracy for each grid pixel.

For each set of $k$ test pixels, construct an error matrix called local error matrix. The overall accuracy derived from each local error matrix is assigned to each grid pixel as local classification accuracy (LCA).

Step 5: Interpolating LCA from grid pixel to the whole map.
The local classification accuracy at grid pixels are interpolated to the whole map using inverse distance weighting (IDW) method as in Foody (2005). The result is what I call map of estimated local classification accuracy.

Let \( u_i = LCA(x_i) \) be the local classification accuracy of grid pixel \( i, i = 1, 2, ..., m \times n \). The local classification accuracy at pixel \( x \) is then estimated as,

\[
u(x) = \sum_{i=1}^{N} w_i(x)u_i \frac{1}{\sum_{j=1}^{N} w_j(x)}
\]

(4.1)

where \( w_i(x) \) is the weight function,

\[
w_i(x) = \frac{1}{d(x, x_i)^p}
\]

(4.2)

and,

- \( x \) denotes target pixel, whose local classification accuracy unknown,
- \( x_i \) are neighbor pixels, whose local classification accuracy are known (estimated using local error matrix),
- \( d \) is the distance between the known pixel \( x_i \) and unknown pixel \( x \),
- \( N \) is total number of known points used in interpolation, \( N \leq m \times n \).
- \( p \) is the power parameter, a positive true number.

The selection of values for \( m, n, \) and \( k \) is subjective. For my data, square grid of size 50 pixels are used. Data 1 and 2 are of size 500×500 pixels, therefore \( m = n = 9 \). There are \( 9 \times 9 = 81 \) grid pixels which will be used to calculate LCA. Note, pixels on the image boundary are not used as grid pixel. Data 3 is of size 1280×307 pixel, I select \( m = 24, n = 5 \), and there are 24×5 grid pixels. For all three datasets, \( k = 150 \). In other words,
for each grid pixel, 150 closest sample pixels from the set of 2000 test samples are used to construct an error matrix. The overall accuracy is derived for each grid pixel.

### 2.2.2 Bootstrap method

Steele et al. (1998) combines bootstrap and kriging to estimate per-pixel misclassification rate. Misclassification rate is estimated at test pixels using bootstrap sampling and then interpolated to the whole image. The process is explained as follows.

**Step 1: Generating bootstrap training set.**

$B$ sets of bootstrap training samples were generated using the original $S_1$ training samples. Each bootstrap set has $S_1$ training samples. Since bootstrap is sampling with replacement, theoretically, the percent of original $S_1$ training samples that will be included in each bootstrap training set is (Hastie et al. 2009)

$$P[(x_i,y_i) \in D | (x_i,y_i) \in C] = 1 - \left(1 - \frac{1}{n}\right)^n \approx 1 - e^{-1} = 0.632.$$

**Step 2: Image classification using bootstrap sample sets.**

The classifier is trained using one bootstrap sample set to compute classification rule, which is then used to classify the image. This training and classification process is applied to each of the $B$ bootstrap sample sets. Each time, I compute a new training rule and then use it to classify the image. In other words, I classify the image $B$ times and get $B$ versions of class maps.

**Step 3: Creating bootstrap class map.**

$B$ class maps are stacked together and each pixel is labeled $B$ times. The mode label for each pixel is recorded and used as the bootstrap class label. This result a bootstrap
class map, which will be further used to create indicator classification confidence map in Section 2.3.1.

Step 4: Calculate misclassification rate (MR) for the test data.

Compare each of the $B$ class maps with the test data and record each test pixel as correctly or incorrectly classified. Suppose test pixel $X_i$ has been incorrectly classified $E_i$ times, misclassification rate is then estimated as $MR = E_i/B$. Note, in the original work of Steele et al. (1998), the term misclassification probability other than misclassification rate is used. I consider misclassification rate a better term.

Step 5: Modeling semivariogram for MR.

In order to kriging, MR is assumed to be an intrinsically stationary spatial process, i.e., the differences of MR between two pixels separated by a given distance have a constant mean and a constant variance. The experimental semivariogram is estimated based on MR of $S_2$ test pixels as,

$$
\gamma(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} [MR(u_{\alpha}) - MR(u_{\alpha} + h)]^2,
$$

where $N(h)$ is the number of pairs of samples with distance $h$ apart from each other. Figure 4.4 shows the experimental semivariogram for MR of data 1 classified using MLC. The semivariogram increase gradually and tend to be stable at a certain lag distance. Mathematical models for fitting experimental semivariogram should satisfy the constraint of positive definiteness. Exponential model of the following form is widely used in the literature (Goovaerts 1997),

$$
\gamma(h) = \begin{cases} 
  a + (\sigma^2 - a)(1-e^{-3h/c}), & \text{for } h > 0 \\
  0, & \text{otherwise}
\end{cases}
$$
where $a$, $\sigma^2$, and $r$ represent nugget effect, sill, and range, respectively. These parameters can be estimated using nonlinear weighted least squares method (Cressie 1985).

![Simivariogram](image)

Figure 4.4 Experimental semivariogram and fitted model for Data 1 classified using MLC

Step 6: Predict MR for the whole map using ordinary kriging.

Under intrinsic stationary assumption for MR, I use ordinary kriging to estimate MR for all the pixels in the class map. The predicted MR for pixel $t$ is a weighted average of the MR for $n$ sample pixels:

$$MR_t = \sum_{i=1}^{n} w_i MR_i$$  \hspace{1cm} (4.5)
where \( w_i \) is the ordinary kriging weight for the \( i \)th sample pixel. The weights are estimated using the semivariogram through solving of the following optimization problem:

\[
\sum_{i=1}^{n} w_i(u) = 1 \\
\min \left\{ w^T(u)Cw(u) + \sigma^2 - 2w^T(u)c(u) \right\}, \quad \text{subject to } w^T(u) \times 1 = 1
\]

where \( C \) is the covariance matrix of MR, \( c \) is the covariance between know pixel and target pixel, \( \sigma^2 \) is the variance of the MR. The ordinary kriging of weight and its variance is then estimated as

\[
w = C^{-1}c \\
\sigma^2_{ok} = \sigma^2 - c^T(u)C^{-1}c(u)
\]

2.2.3 Geostatistical method

Different from the previous two interpolation based methods, Kyriakidis and Dungan (2001) propose a geostatistical method to map per-pixel classification confidence. Their central idea is to combine error matrix with kriging to predict per-pixel classification confidence. The geostatistical method includes the following steps.

Step 1: Image classification.

Classify image in the usual way using \( S_1 \) training data, same as Step 1 for local error matrix method. Construct the regular (global) error matrix using \( S_2 \) test data. For the convenience of discussion bellow, a typical error matrix is shown in Table 4.1. The class map will also be used to create indicator classification confidence map in Section 2.3.1.

Step 2: Define indicator variable of class labels for each pixel.
Let the class label of pixel \( u \) be a random variable \( s(u) \). Based on the indicator framework of Journel (1986), class label for each pixel can be coded as a set of \( K \) local probabilities, each associated with the \( k \)th class \( k \):

\[
Pr\{s(u) = k \mid \text{info}(u)\}s,k = 1, 2, ..., K, \tag{4.8}
\]

which represents the probability of class \( k \) observed at location \( u \) on the ground given the classification results \( \text{info}(u) \).

Table 4.1 A typical error matrix

<table>
<thead>
<tr>
<th>Class map</th>
<th>Reference data</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>( x_{11} )</td>
<td>( x_{12} )</td>
</tr>
<tr>
<td>2</td>
<td>( x_{21} )</td>
<td>( x_{22} )</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>( k' )</td>
<td>( x_{k'1} )</td>
<td>( x_{k'2} )</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>( K )</td>
<td>( x_{K1} )</td>
<td>( x_{K2} )</td>
</tr>
<tr>
<td>Column total</td>
<td>( x_{+1} )</td>
<td>( x_{+2} )</td>
</tr>
</tbody>
</table>
Equation (2.4) can be further specified based on the different information used (Goovaerts 1997). For image classification, the following two scenarios are used.

1. For the reference map, equation (2.4) turns out as follows,

\[
I(u;k) = \begin{cases} 
1, & \text{if } s(u) = k \\
0, & \text{if not} \\
\end{cases} \quad k = 1, 2, ..., K. \tag{4.9}
\]

2. For classification map, equation (2.4) is expressed as,

\[
y(u;k) = Pr\{s(u) = k | x(u) = k'\} = p(k | k'), k = 1, 2, ..., K, \tag{4.10}
\]

where \(p(k | k')\) is the proportion of a pixel of class \(k\) on the reference map, given that it is classified as class \(k'\), and \(x(u)\), the class label for pixel \(u\) on the class map. \(p(k | k')\) can be estimated as:

\[
p(k | k') = \frac{\sum_{u=1}^{n(u)} I(u_a; k) J(u_a; k')}{\sum_{a=1}^{n(u)} J(u_a; k')}, \quad k, k' = 1, 2, ..., K, \tag{4.11}
\]

where \(I(u_a; k)\) is the indicator for reference map \(s(u; k)\) as in equation (2.6), \(J(u_a; k')\), the indicator of classification map \(x(u; k')\) defined as

\[
J(u; k') = \begin{cases} 
1, & \text{if } x(u) = k' \\
0, & \text{if not} \\
\end{cases} \quad k' = 1, 2, ..., K. \tag{4.12}
\]

In practice, \(p(k | k')\) is estimated using the error matrix as in Table 4.1, i.e., divide \(x_{k'k}\) by corresponding row total \(x_{k'\cdot}\) and obtain

\[
p(k | k') = \frac{x_{k'k}}{x_{k'\cdot}} = \frac{x_{k'k}}{\sum_{k=1}^{K} x_{k'k}}, \tag{4.13}
\]

Specifically, when \(k' = k\), \(p(k | k')\) is the user’s accuracy.

Step 3: Define residual for class probability.
The conditional probability in equations (4.11) and (4.13) can be viewed as the mean of the indicator RV \( I(u;k) \) at location \( u \):

\[
p(k | k') = E \{ I(u;k) \}, k = 1, 2, ..., K. \tag{4.14}
\]

In other words, the conditional probability obtained from the error matrix is the "average" spatial variability of land classes on the ground. Therefore, the residual for class probabilities can be defined as follows,

\[
r(u_a;k) = I(u_a;k) - y(u_a;k). \tag{4.15}
\]

Step 4: Model the residuals using empirical semivariogram similar to step 5 in Section 2.2.2. As usual, exponential model is used to fit the semivariogram.

Step 5: Predict probabilities of class \( k \) allocated to pixel \( u \) using simple indicator kriging:

\[
p_{stk}^*(u;k) = y(u;k) + \sum_{a=1}^{n(u)} w(u_a;k)[I(u_a;k) - y(u_a;k)], \tag{4.16}
\]

where the weights \( w(u_a;k) \) are determined by solving the simple kriging system:

\[
\min \{ w^T(u;k)Cw(u;k) + \sigma(k)^2 - 2w^T(u;k)e(u;k) \} \tag{4.17}
\]

where \( C \) is the covariance matrix of residual \( r(u;k) \), \( e \) is the covariance between know pixel and target pixel, \( \sigma^2 \) is the variance of the \( r(u;k) \). The simple kriging of weight and its variance is then estimated as

\[
w(u;k) = C^{-1}e(u;k)
\]

\[
\sigma(k)_{ok}^2 = \sigma(k)^2 - e^T(u;k)C^{-1}e(u;k) \tag{4.18}
\]

Step 6: Construct per pixel confusion index based on \( p(u;k) \).
According to suggestions of Kyriakidis and Dungan (2001), the following local index of map quality, called per pixel confusion index (CI), is constructed,

\[
c(u) = \left[1 - p^m(u)\right] \left(\frac{K}{K - 1}\right),
\]

where \(p^m(u) = \max\{p(u; k), k = 1, 2, \ldots, K\}\), and \(K/(K-1)\) is a standardization factor. When a pixel is classified with without uncertainty, i.e., \(p^m(u) = 1\), \(c(u) = 0\). When the probability of a pixel belonging to each class is equal, the lowest classification confidence is arrived. In such case, \(p^m(u) = p(u; k) = 1/K\), \(c(u) = 1\).

### 2.2.4 Clarification of concepts

It is necessary to clarify four concepts used in this study: local classification accuracy (LCA), misclassification rate (MR), confusion index (CI), and per-pixel classification confidence.

- **Local classification accuracy (LCA)** is used by Foody (2005), referring to the overall accuracy derived from local error matrix. The higher the LCA is, the more accurate the classification is.

- **Misclassification rate (MR)** is used by Steele et al. (1998) to measure the number of misclassification divided by the total number of bootstrap classification. The lower MR is, the less uncertain the classification is.

- **Confusion index (CI)** is defined by Kyriakidis and Dungan (2001) to characterize the confusion of class assignment in image classification. The lower CI is, the less confusion of class assignment, i.e., the less uncertain the classification is.
Per-pixel classification confidence is defined as the probability of a pixel being correctly classified.

2.3 Estimated classification confidence vs. true classification confidence

In order to evaluate the performance of three methods, I compare the map of estimated classification confidence with a map of true classification confidence. Using each of the three methods in Section 2.2.1-2.2.3, I get a map of estimated per-pixel classification confidence, which takes the form of continuous values ranging between 0-1. For convenience, I call the estimated classification confidence map as the continuous map of estimated classification confidence. Local error matrix method and bootstrap method have similar procedures to estimate per-pixel classification confidence: (1) Generate classification confidence at sample pixels; (2) Interpolate estimation at sample pixels to the whole map. Different from these two methods, the geostatistical method generates a set of class probabilities for each pixel belonging to each class. These class probabilities are then used to create a local index of classification confidence.

Due to the difference in three methods, I design different approaches to obtain the map of true classification confidence.

First, for each of the three methods, a class map is obtained through the methods in 2.2.1-2.2.3. The class map is then compared with the full coverage reference map to create an indicator map where each pixel is indicated as either correctly classified or incorrectly classified.

Second, for local error matrix method and bootstrap method, I design another type of true classification map in continuous form. I create this continuous map of classification confidence by extending the original method to all pixels without
interpolation. Procedures for this second type of true classification confidence map are different for local error matrix method and bootstrap method. This continuous true classification confidence map is used to examine the interpolation effect of both methods.

For local error matrix method, I obtain a continuous map of true classification confidence by extending the local error matrix from sample pixels to all the pixels. In other words, I construct local error matrix as Step 3 and 4 in Section 2.2.1 for each pixel, not just for the grid pixel. My full coverage reference data make this possible. The overall accuracy estimated from each local error matrix is regarded as the true classification confidence for each pixel.

For bootstrap method, similar to Step 4 in Section 2.2.2, I create a map of true MR by comparing each bootstrap class map with the full coverage map and calculate MR for each pixel, not just for the test sample pixels.

It should be noted that due to the special procedures of geostatistical method, there is no way to obtain a meaningful map of true classification confidence in continuous form.

2.4 Approaches to evaluating method performance

2.4.1 Evaluation scheme

Table 4.2 shows the evaluation scheme for this study.

First, I evaluate two relationships between the classification confidence map derived from each of the tree methods. For all three methods, I divide the pixels in each datasets into two groups: correct pixels and error pixels. I then examine the relationship between the estimated classification confidence for correct pixels and error pixels. It is assumed that correct pixels tend to have high classification confidence, while error pixels
have low classification confidence. In other words, the classification confidence should be distinguishable for correct pixels and error pixels. For local error matrix method and bootstrap method, I will also examine the relationship between the estimated continuous classification confidence and true continuous classification confidence. If an interpolation based method is effective in predicating classification confidence, the estimated classification confidence should agree well with the true classification confidence.

Second, I evaluate the three methods in two ways: exploratory data analysis (EDA) and statistical analysis. For exploratory data analysis, I use one EDA tool, bi-histogram. For statistical analysis, I construct two statistics, Willmott's $d$ and Binned classification quality (BCQ). Willmott's $d$ measures the similarity between the estimated continuous classification confidence map and the true continuous classification confidence map (Willmott 1981, 1982; Willmott et al. 2012). BCQ is calculated based on binned pixels and thus examines the relationship between LAC/MR/CI and proportion of correct pixels in each bin.

Table 4.2 Evaluation scheme

<table>
<thead>
<tr>
<th></th>
<th>EDA</th>
<th>Statistical analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated classification confidence for correct and error pixels</td>
<td>LEM, BM, GM</td>
<td>LEM, BM, GM</td>
</tr>
<tr>
<td>Continuous maps of estimated vs. true classification confidence</td>
<td>LEM, BM</td>
<td>LEM, BM</td>
</tr>
</tbody>
</table>

Note: LEM - local error matrix method, BM - bootstrap method, GM - geostatistical method.


2.4.2 Bi-histogram

As stated in Chapter 3, the bi-histogram is a graphical tool for examining the distribution of two datasets by the histograms of both datasets. In this chapter, I use bi-histogram to explore the two types of relationships introduced in Table 4.2. I will plot two bi-histograms: (1) bi-histogram of the estimated classification confidence of correct pixels vs. error pixels; (2) bi-histogram of the estimated vs. true continuous classification confidence (for local error matrix method and bootstrap method only).

2.4.3 Measure of agreement

As discussed in Section 2.3, for local error matrix method and bootstrap method, continuous maps of true and estimated classification confidence are obtained. In this section, the agreement of two maps is evaluated. If both maps agree well with each other, the method is effective in estimating per-pixel classification confidence.

There are several widely used measures of agreement/disagreement of two datasets (Ji and Gallo 2006), including Pearson correlation coefficient ($r$), coefficient of determination ($r^2$), mean absolute error (MAE), root mean square error (RMSE), Willmott's index of agreement ($d$), Mielke's measure of agreement ($\rho$), Robinson's coefficient of agreement ($A$), and Ji and Gallo's agreement coefficient ($AC$). A detailed review of these measures is referred to Ji and Gallo (2006). In this chapter, I use Willmott's $d$ to measure the agreement between the continuous map of estimated classification confidence and true classification confidence. Willmott's $d$ has two main advantages that serve my purpose: (1) Bounded in a range from 0 to 1, indicating degree of agreement ranging from complete disagreement to complete agreement. (2) Non-dimensional, thus easier to interpret than the widely used RMSE.
Similar to most dimensionless measures of agreement, Willmott's $d$ is designed in the following form (Willmott 1981, 1982; Willmott et al. 2012)

$$\rho = 1 - \frac{\delta}{\mu},$$

(4.20)

where $\delta$ is a dimensioned average error-magnitude, $\mu$ the potential error, i.e., the basis of comparison. Willmott's $d$ is expressed as

$$d = 1 - \frac{SSE}{PE}$$

$$= 1 - \frac{\sum_{i=1}^{N} [(X_i - \bar{X}) - (Y_i - \bar{X})]^2}{\sum_{i=1}^{N} |X_i - \bar{X}| - |Y_i - \bar{X}|^2}$$

(4.21)

$$= 1 - \frac{\sum_{i=1}^{N} (X_i - Y_i)^2}{\sum_{i=1}^{N} |X_i - \bar{X}| - |Y_i - \bar{X}|^2}$$

where $X_i$ denotes true classification confidence, $\bar{X}$ is the mean of $X_i$, $Y_i$ is the predicated classification confidence, $N$ is the number of pixels in the map. Willmott's $d$ is a measure based on the sum of squares, $\delta$ is the sum of the squared errors, and $\mu$ is the overall sum of the squares for absolute values of two partial differences from the true mean, $|X_i - \bar{X}|$ and $|Y_i - \bar{X}|$. According to equation (4.21), the lower limit of $d$ is zero, indicating complete disagreement, and the upper limit of $d$ is one, indicating complete agreement.

### 2.4.4 Binned classification quality

Using the three methods introduced in Section 2.2.1-2.2.3, three measures of per-pixel classification confidence are obtained: LCA, MR, and CI. Similar to Section 2.4.2 in Chapter 3, binned classification quality (BCQ) is calculated using LCA, MR, and CI.
After calculation of BCQ for LCA, MR, and CI, the following hypotheses are examined:

- For LCA, BCQ$_q$ has positive relationship with $b_q$.
- For MR, BCQ$_q$ has negative relationship with $b_q$.
- For CI, BCQ$_q$ has negative relationship with $b_q$.

### 2.5 Summary of methodology

Here is a summary of the methodology for this chapter.

Step 1: Image classification. The three datasets are classified using four commonly used classifiers: MLC, NN, SVM, and BDT. For local error matrix method and geostatistical method, each involves $3 \times 4 = 12$ scenarios of image classification and thus results 12 versions of class maps. For bootstrap method, there are $12 \times B$ versions of class maps.

Step 2: Creating classification confidence map. For local error matrix method and bootstrap method, two types of classification confidence map are obtained: continuous map and indicator map. For geostatistical method, only the indicator classification confidence map is generated.

Step 3: Calculating two statistics: Willmott's $d$ and BCQ

- Willmott's $d$ is calculated by comparing estimated LCA/MR with the true LCA/MR respectively.
- BCQ is calculated by grouping pixels based on each of the three measures, i.e., LCA, MR, and CI.
Step 4: Evaluating three methods using EDA tool (bi-histogram) and two statistics (BCQ and Willmott's $d$).

The results from three methods are also compared to examine their difference in predicting per-pixel classification confidence.

3 Results

3.1 Local error matrix method

3.1.1 Maps of classification confidence

Figure 4.5 shows the maps of estimated LCA based on maximum likelihood classification. Each panel is created by interpolating the overall accuracy from the local error matrix for each grid pixel. There are some spatial patterns in the estimated LCA map. For example, Data 1 has low LCA in west and southwest side, and a bump of high LCA stretching from southwest to northeast. According to the reference map of Data 1, those areas with low LCA are highly heterogeneous patches, while areas with high LCA are homogeneous agricultural land. Data 2 has a valley of low LCA stretching from southwest to northeast, a hotspot of high LCA in southeast corner, and a high patch in the north side. Areas with low LCA are mainly composed of residential building and concrete surfaces, and these two classes are easily to be confused. Areas with high LCA are mainly forest areas and are more homogeneous. Data 3 has low LCA stretching from mid-north to mid-south and southeast, and high LCA in mid-west and northeast. Similar to Data 2, areas of low LCA in Data 3 is due to the confusion between roof and concrete surfaces. Areas of high LCA are vegetated areas which are of less confusion.
Figure 4.5 Map of estimated LCA from IDW interpolation (MLC)

(a) - Data 1
(b) - Data 2
(c) - Data 3

Figure 4.6 Maps of true LCA (MLC)
Figure 4.6 shows the true LCA based on maximum likelihood classification which is the overall accuracy from the local error matrix for each pixel. Figure 4.7 shows the indicator maps of classification error/correct which are generated by comparing the class map with the full coverage reference map. As expected, plots in Figure 4.6 are smoother than the plots in Figure 4.7, while plots in Figure 4.5 are much smoother than those in Figure 4.6. The visual difference of three figures is quite evident.

### 3.1.2 Estimated LCA of correct and error pixels

Figure 4.8 shows the bi-histogram of estimated LCA for correct pixels and error pixels. The shapes of corresponding histograms for both types of pixels look quite similar. Most pixels, either correctly classified or incorrectly classified, have LCA
concentrated in certain range. For Data 2, correct pixels seem to be a little more concentrated in high LCA than error pixels.

![Figure 4.8 Bi-histogram of estimated LCA for correct and error pixels](image)

Note: y - LCA; x: proportion of pixels. Left side - correct pixel, right side - error pixel.

### 3.1.3 Comparing the continuous map of estimated and true LCA

The visual difference between Figure 4.5, Figure 4.6 is evident. First, the distribution of LCA is different. Figure 4.9 gives the bi-histograms of true vs. estimated LCA for all three data with four classifiers. The true LCA is more spread out than the estimated LCA.
Figure 4.9 Bi-histogram of true and estimated LCA

Note: y: LCA, x: proportion of pixels. Left - true LCA, right - estimated LCA.

Table 4.3 Willmott's $d$ of estimated and true LCA

<table>
<thead>
<tr>
<th></th>
<th>MLC</th>
<th>NN</th>
<th>SVM</th>
<th>BDT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data 1</strong></td>
<td>0.224</td>
<td>0.262</td>
<td>0.249</td>
<td>0.220</td>
</tr>
<tr>
<td><strong>Data 2</strong></td>
<td>0.449</td>
<td>0.401</td>
<td>0.393</td>
<td>0.430</td>
</tr>
<tr>
<td><strong>Data 3</strong></td>
<td>0.291</td>
<td>0.260</td>
<td>0.242</td>
<td>0.283</td>
</tr>
</tbody>
</table>
In summary, local error matrix method produces poor results for classification confidence. This is due to the internal weakness of local error matrix method, i.e., the use of interpolation.

3.1.4 Relationship between binned classification quality and LCA

As introduced in Section 2.4.4, the hypothesis to test is: If a pixel has high LCA, the probability of this pixel being a correct pixel is high. In other words, there should be a positive relationship between BCQ\(_q\) and \(b_q\). Figure 4.10 shows scatter plot of binned classification quality (BCQ\(_q\)) against LCA (\(b_q\)) where LCA is divided into 30 bins. The patterns of the scatter plots vary with data and classifiers. For Data 1, there seems no clear relationship between BCQ\(_q\) and \(b_q\). The scatter plots of Data 2 show linear positive relationship between BCQ\(_q\) and \(b_q\) although there are some outliers when LCA is small. The scatter plots of Data 3 show some linear positive relationship between BCQ\(_q\) and \(b_q\) with MLC. As to Data 3 with NN, SVM, BDT, the scatter plots show a "V" shape pattern. Besides dividing LCA into \(Q = 30\) bins, I also tried other values of \(Q = 40, 50, ..., 100\). The general pattern of the scatter plot is not sensitive to the number of bins, \(Q\).

Table 3.2 shows the correlation coefficient (R) between BCQ\(_q\) and \(b_q\). The low values of R indicate weak correlation between BCQ\(_q\) and \(b_q\). It can be concluded that LCA is not a good measure for characterizing per-pixel classification confidence.
Figure 4.10 Scatter plot of BCQ<sub>q</sub> vs. \( b_q \) based on LCA for all the pixels

Table 4.4 Correlation coefficients (R) of BCQ<sub>q</sub> vs. \( b_q \) for LCA (\( Q=30 \))

<table>
<thead>
<tr>
<th></th>
<th>MLC</th>
<th>NN</th>
<th>SVM</th>
<th>BDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data 1</td>
<td>0.5349</td>
<td>0.6569</td>
<td>0.2932</td>
<td>0.6547</td>
</tr>
<tr>
<td>Data 2</td>
<td>0.8291</td>
<td>0.8309</td>
<td>0.8362</td>
<td>0.9634</td>
</tr>
<tr>
<td>Data 3</td>
<td>0.8821</td>
<td>0.8350</td>
<td>0.8861</td>
<td>0.5532</td>
</tr>
</tbody>
</table>
3.2 Bootstrap method

3.2.1 Maps of classification confidence

Figure 4.11 shows the maps of estimated MR based on maximum likelihood classification. Each panel is created by interpolating the MR from $S_2$ test sample pixels using ordinary kriging. There are some spatial clusters in the MR map. The clusters in Data 1 seem to spread randomly across the whole map, while the clusters for Data 2 and 3 have a certain spatial pattern: high MR clusters are mainly located in the roof and concrete surface area. This confirms previous findings in Section 3.1 that classifiers have higher error in distinguishing roof and concrete surfaces.

Figure 4.12 shows the true MR based on bootstrap method. MR in each panel is obtained by comparing each bootstrap class map with the full coverage reference map and record each pixel as correctly or incorrectly classified. Same as MR for test pixels, MR for each pixel is the number of misclassification times divided by the number of bootstrap iteration, i.e., $B$. Figure 4.13 shows the indicator maps of classification error generated by comparing the bootstrap class map with the full coverage reference map. Figure 4.11 is visually different from both Figure 4.12 and Figure 4.13 in that the interpolation smoothed out true classification confidence. Therefore, bootstrap method may not be a good method for per-pixel classification confidence. Again, this is due to the internal weakness of using interpolation to estimate classification confidence.
Figure 4.11 Maps of estimated MR from bootstrap method (MLC)

Figure 4.12 Maps of true MR from bootstrap method (MLC)
3.2.2 Estimated MR of correct and error pixels

Same as the evaluation of local error matrix method, the bi-histogram of MR for correct pixels and error pixels are plotted in Figure 4.14. MR for correct and error pixels have similar distribution. The visual shapes of corresponding histograms for correct pixels and error pixels look quite similar except for Data 2. Most pixels, either correctly classified or incorrectly classified, have MR concentrated below 0.3. The similar distribution of correct pixels and error pixels indicate that MR cannot be used to distinguish both types of pixels. It may be concluded that MR is not a good measure for per-pixel classification confidence.
Figure 4.14 Bi-histogram of MR for correct and error pixels

Note: y - MR; x: proportion of pixels. Left side - correct pixel, right side - error pixel.

3.2.3 Comparing the continuous map of estimated and true MR

Figure 4.15 gives the bi-histogram of estimated and true MR for all three data with four classifiers. The difference of true and estimated MR is clear. True MR is mainly concentrated in low values, while the estimated MR is more spread out, especially for Data 2 and Data 3.

Table 4.5 shows the Willmott's $d$ calculated from the true and estimated MR, indicating low agreement between estimated MR and true MR. The result from Willmott's $d$ confirms the findings in Figure 4.15.
Figure 4.15 Bi-histogram of true and estimated MR

Note: y: MR, x: proportion of pixels. Left - true values, right - estimated values.

Table 4.5 Willmott's $d$ for estimated and true MR

<table>
<thead>
<tr>
<th></th>
<th>MLC</th>
<th>NN</th>
<th>SVM</th>
<th>BDT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data 1</strong></td>
<td>0.2656</td>
<td>0.2023</td>
<td>0.3852</td>
<td>0.2047</td>
</tr>
<tr>
<td><strong>Data 2</strong></td>
<td>0.5517</td>
<td>0.5444</td>
<td>0.6162</td>
<td>0.5858</td>
</tr>
<tr>
<td><strong>Data 3</strong></td>
<td>0.4729</td>
<td>0.4072</td>
<td>0.4393</td>
<td>0.4480</td>
</tr>
</tbody>
</table>

In summary, bootstrap method is not efficient for estimating classification confidence. The key issue is due to the interpolation effect.
3.2.4 Relationship between binned classification quality and MR

Figure 4.16 shows the scatter plots of binned classification quality (BCQ,) against MR (b,) where MR is divided into 30 bins with equal distance. There is negative relationship between BCQ, and b,. The higher MR is, the higher probability that a pixel is an error pixel. Figure 4.16 also shows that the exact relationships between different datasets and classifiers are different. For Data 1 with MLC, NN, and BDT, the curves are steep when MR< 0.8, 0.7, 0.6 respectively, and then they level off. For Data 1 with SVM, Data 2 with all classifiers, the curves look very similar: BCQ decreases slowly when MR<0.8, then BCQ drops quickly to zero. For Data 3, the curves of the scatter plots decrease slowly when MR is smaller than 0.9, and then drop sharply when MR is greater than 0.9.

It should be noted that there are some level off points in the scatter plots for Data 1 with MLC, NN, BDT. This indicates that pixels with MR greater than certain cutoff values are all misclassified. The cutoff value for Data 1 with MLC, NN, and BDT are 0.9, 0.7667, and 0.7333.

Table 4.6 shows the correlation coefficient R between BCQ,q and b,q. The absolute values of R are not high which indicate the correlation between BCQ,q and b,q is not strong. It can be concluded that LCA is not a good measure for characterizing per-pixel classification confidence.
Figure 4.16 Scatter plot of BCQ$_q$ vs. $b_q$ based on estimated MR

Table 4.6 Correlation coefficients (R) of BCQ$_q$ vs. $b_q$ for MR ($Q=30$)

<table>
<thead>
<tr>
<th></th>
<th>MLC</th>
<th>NN</th>
<th>SVM</th>
<th>BDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data 1</td>
<td>-0.9525</td>
<td>-0.9551</td>
<td>-0.8792</td>
<td>-0.9539</td>
</tr>
<tr>
<td>Data 2</td>
<td>-0.9917</td>
<td>-0.9943</td>
<td>-0.9977</td>
<td>-0.9945</td>
</tr>
<tr>
<td>Data 3</td>
<td>-0.8886</td>
<td>-0.8889</td>
<td>-0.8087</td>
<td>-0.8628</td>
</tr>
</tbody>
</table>
3.3 Geostatistical method

3.3.1 Maps of estimated classification confidence

Figure 4.17 shows the maps of estimated CI based on maximum likelihood classification. Each panel is created using the geostatistical method introduced in Section 2.2.3. There are some spatial patterns in the estimated CI map. For example, data 1 has some hotspots scattered across the map. For Data 2 and 3, confusion is high in roof and concrete surface area. Different from the other two methods, the geostatistical method does not provide a true CI map. The indicator map of true classification error is the same as that for the local error matrix method, which is shown in Figure 4.7. Figure 4.17 and Figure 4.7 show that CI estimated from geostatistical method have quite different spatial patterns of true classification error.

![Map of estimated CI (MLC)](image)

Figure 4.17 Map of estimated CI (MLC)
3.3.2 Estimated CI of correct and error pixels

Figure 4.18 shows the bi-histogram of CI for correct pixels and error pixels. CI for correct pixels tends to concentrate at lower values, while CI for error pixels are more spread out across the range of [0,1] except for Data 1. For Data 1, CI for correct pixels concentrated between 0-0.2 while CI for error pixels have another small peak at the intervals of [0.3,0.5]. For Data 2, CI for correct pixels concentrated in [0-0.2] while CI for error pixels quite spread out on [0, 0.6]. CI for Data 3 have more variability with correct pixels spread out in [0, 0.5] and error pixels in [0, 0.7].

Figure 4.18 Bi-histogram of CI

Note: y - CI; x: proportion of pixels. Left side - correct pixel, right side - error pixel.
3.3.4 Relationship between binned classification quality and CI

Figure 4.19 shows the scatter plots of binned classification quality ($BCQ_q$) against CI ($b_q$) where CI is divided into 30 bins with equal distance. Generally, there is a negative relationship between $BCQ_q$ and $b_q$. However, the trend is not monotonic, especially for Data 2 with NN and SVM, Data 1 and Data 3 for all classifiers. The scatter plots for Data 2 with MLC and BDT mostly have a monotonically decreasing trend. In summary, CI is not a good measure for predicating per-pixel classification confidence.

![Figure 4.19 Scatter plot of BCQ_q vs. b_q based on CI](image)
Table 4.7 shows the correlation coefficient R between BCQ_q and b_q. The absolute values of R vary great with datasets and classifiers. They are high for Dat 1 with MLC, NN, and SVM, Data 2 with all classifiers, and Data 3 with MLC, and BDT, and low in other scenarios.

It can be concluded that LCA is not a good measure for characterizing per-pixel classification confidence.

Table 4.7 Correlation coefficients (R) of BCQ_q vs. b_q for CI (Q=30)

<table>
<thead>
<tr>
<th></th>
<th>MLC</th>
<th>NN</th>
<th>SVM</th>
<th>BDT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data 1</strong></td>
<td>-0.9660</td>
<td>-0.9370</td>
<td>-0.9215</td>
<td>-0.7857</td>
</tr>
<tr>
<td><strong>Data 2</strong></td>
<td>-0.9868</td>
<td>-0.9656</td>
<td>-0.9555</td>
<td>-0.9856</td>
</tr>
<tr>
<td><strong>Data 3</strong></td>
<td>-0.9774</td>
<td>-0.7181</td>
<td>-0.8857</td>
<td>-0.9286</td>
</tr>
</tbody>
</table>

### 4 Discussion and conclusions

The performance of LCA, MR, and CI varies across data and classifiers. Table 4.8 is a summary of the results for three interpolation based methods. The results can be summarized as follows.

(1) Local error matrix method is the least reliable one among the three methods. There is significant difference between estimated LCA and the true LCA. The estimated LCA are mostly concentrated around the overall accuracy of the whole map (between 0.7-0.9) while the true LCA are much more spread. The distribution of correct and error pixels are similar, which means LCA cannot distinguish error pixels from correct pixels.
In other words, local error matrix is not a good method for estimating per-pixel classification confidence.

Table 4.8 Comparing the results of three interpolation based methods

<table>
<thead>
<tr>
<th>Items</th>
<th>Local error matrix method</th>
<th>Bootstrap method</th>
<th>Geostatistical method</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification confidence of correct and error pixels</strong></td>
<td>Bi-histogram shows that the distribution of correct and error pixels are similar.</td>
<td>Bi-histogram shows that the distribution of correct and error pixels are similar.</td>
<td>Bi-histogram shows that correct pixels tend to concentrate at lower values, while error pixels are more spread.</td>
</tr>
<tr>
<td><strong>Estimated and true classification confidence</strong></td>
<td>Estimated LCA are more spread than true LCA.</td>
<td>True MR for are mostly concentrated in low values while estimated MR are more spread out.</td>
<td>N.A.</td>
</tr>
<tr>
<td><strong>Binned classification quality</strong></td>
<td>There is no clear relationship between BCQ and LCA.</td>
<td>BCQ are negatively related to MR.</td>
<td>BCQ is somewhat negatively related to CI. However, the relationship is not monotonic for many scenarios.</td>
</tr>
</tbody>
</table>

(2) Bootstrap method may be useful in estimating per-pixel classification confidence. However, the distributions of estimated MR for correct and error pixels are very similar which means MR is not a good measure to distinguish error pixels from correct pixels. Also, the estimated MR is different from the true MR. There does exist correlation between BCQ and MR. BCQ is generally negatively related to MR but the exact relationship between binned classification quality and MR varies with datasets and classifiers.
(3) Geostatistical method is of limited use and is not reliable in estimating per-pixel classification confidence. The distributions of correct pixels and error pixels are different. CI of correct pixels tends to concentrate at lower values, while error pixels are more spread. While there is generally negative relationship between binned classification quality and CI, the relationship is not monotonic for some datasets and classifiers. Pixels with high CI may also have high classification confidence.

The three interpolation based methods and their corresponding indices characterize the per-pixel classification confidence in different ways.

Local error matrix method is in essence a kind of smoothing window method. The LCA for each grid pixel is the unweighted average classification accuracy estimated based on its neighboring test pixels through the tool of local error matrix. The LCA for non-grid pixel is the weighted average based on neighboring grid pixels through IDW interpolation. Besides IDW, other interpolation techniques, e.g., kriging, can also be used. Due to the properties of local error matrix method, map of LCA is very smooth. My study shows that it is least useful in predicting per-pixel classification confidence.

Bootstrap method characterizes the separability of a pixel among different classes by resampling training data many times and thus results different classification rules. If a pixel cannot be classified correctly using different classification rules, then it is highly probable that this pixel will be classified wrong in practice. This study shows that bootstrap method is of limited use. Although there is negative relationship between BCQ and MR, the exact mathematical relationship varies with datasets and classifiers. As an interpolation method, the estimated MR map is smooth compared to the true MR map. Another issue is that MR cannot distinguish error pixels from correct pixels well.
Geostatistical method combines global error matrix with local variation of assigning class label to each pixel. Geostatistical method generates conditional probabilities, $p(u;k)$ of class labels for each pixel. These conditional probabilities are similar to the posterior probabilities, $p$, directly from classifiers. The difference between $p(u;k)$ and $p$ is: $p(u;k)$ use the information of test data while $p$ does not. The geostatistical method provides ideas for further studies. For example, it may be used to improve image classification. Geostatistical method is not reliable to predict per-pixel classification confidence either due to the fact that the relationship between BCQ and CI is not monotonic.

In conclusion, the three interpolation methods provide some interesting insights on various aspects of estimating per-pixel classification confidence. Unfortunately, the interpolation assumes that classification confidence is smooth across the space. However, this is usually not true in practice. The interpolation effects hinders their practical use.
1 Introduction

The valuation in Chapter 3 and 4 shows that classification score based method is promising estimating per-pixel classification confidence. It has also been found that margin of victory derived from probability vectors is most consistently related to per-pixel classification confidence. I also suggest that a transformation of MV is necessary to provide a better estimate of classification quality.

In this chapter, I propose a regression method to convert MV into classification confidence by incorporating information from a separate test dataset. The method is evaluated with full coverage reference data. I experiment my method on three representative datasets with four commonly used classifiers: maximum likelihood classifier (MLC), neural network (NN), support vector machine (SVM), and boosted decision tree (BDT). The rest of the paper is organized as follows. Section 2 is the methodology part. Section 3 is the results, followed by discussion and conclusions in Section 4.

2 Methodology

The methodology of this chapter is illustrated in Figure 3.1 which is composed of four parts.
Part A is about data: (1) remote sensing image, (2) reference data - the selection of training data and test data. Part B is image classification: (1) classifying image to produce probability vectors and class map; (2) calculating margin of victory (MV); (3) compare class map with test data to create indicator map of classification correct/error. Part C is to estimate per-pixel classification confidence: (1) calculating binned classification quality; (2) examining the relationship between BSQ and MV; (3) estimating per-pixel classification confidence (PPCC). Part D is method evaluation. I evaluate the method in
four aspects: (1) exploratory data analysis (EDA) using bi-histogram of PPCC for correct and error pixels; (2) statistical relationship between binned PPCC (BPC) with PPCC using full coverage reference data; (3) stability of PPCC estimates; (4) impact of sample design on PPCC estimates.

2.1 Test datasets

The test datasets used in this chapter are the same as those introduced in Section 2.1 of Chapter 3. Please refer to it accordingly.

For each dataset, randomly select $S_1 + S_2$ sample pixels, where $S_1$ pixels are used as training set, and the remaining $S_2$ pixels as test set. $S_1$ and $S_2$ is non-overlapped and independent with each other. In this study, I set $S_1 = S_2 = N = 2000$.

2.2 Image classification

2.2.1 Probability vectors and margin of victory

The three images are classified in the ordinary way to produce probability vectors from each classification process. The classification scores used in this study are the class probabilities output from each classifier. Each of the three datasets will be classified using four commonly used classifiers: maximum likelihood classifier (MLC), neural network (NN), support vector machine (SVM), boosted decision tree (BDT).

Suppose remote sensing image $X$ with $d$ bands is classified into $K$ classes. There are $S$ pixels in total. Each classifier produces a probability vector $\{p_k = P(Y = \omega_k | X)\}$ for each pixel, where $X = (x_1, x_2, \ldots, x_d)$, represents the spectral data of the pixel, $X \in \mathcal{X}$, $\omega_k$ is the class label, $k = 1, 2, \ldots, K$, and $K$ is the number of classes. The training data are
\( Tr = \{(X_1,Y_1),..., (X_N,Y_N)\} \), where \( X_i \in \mathcal{X} \), \( w_1,...,w_k \), \( i = 1, 2, ..., N \). The classification is done by assigning each pixel \( X \) to the most probable class, i.e.,

\[
f(X) = \omega_k \quad \text{if} \quad P(\omega_k | X) > P(\omega_l | X) \quad \text{for all} \quad l \neq k.
\]

(5.1)

According to McIver and Friedl (2001), MV is the difference in the class probabilities between the most probable class and the second most probable class:

\[
MV = p_1 - p_2.
\]

(5.2)

MV ranges between 0 and 1. The extrema of MV arrives in the following conditions:

- \( MV_{\text{min}} = 0 \), when \( p_1 = p_2 \).
- \( MV_{\text{max}} = 1 \), when \( p_1 = 1 \) and \( p_k = 0 \) for all \( 1 < k \leq K' \).

### 2.2.2 Indicator variable of classification correct/error

Define the indicator variable of classification correct/error by comparing class map with \( S_2 \) test data,

\[
I(u) = \begin{cases} 
1, & \text{if } x(u) = s(u), \\
0, & \text{otherwise},
\end{cases}
\]

(5.3)

where \( x(u) \) is the class label of pixel \( u \) assigned by the classifier, \( s(u) \) is the class label of pixel \( u \) on the reference map. Using this indicator variable, I group the pixels at test locations into correct pixels and error pixels.

### 2.3 Estimate per-pixel classification confidence

#### 2.3.1 Binned classification quality

The procedures to estimate Binned classification quality is the same as that presented in Section 2.4.2 of Chapter 3. For complete development of the method, it is
repeated as follows. Here, BCQ calculated using margin of victory based on the $S_2$ test pixels.

Step 1: Grouping pixels into bins.

The $S_2$ test pixels are divided into a set of $Q$ bins based on the values of MV. Let $M_{\min}$ and $M_{\max}$ be the minimum and maximum values for MV, the range $[M_{\min}, M_{\max}]$ is divided into $Q$ intervals:

$$M_q = \begin{cases} 
(b_{q-1}, b_q], & \text{when } q = 2, 3, \ldots, Q, \\
[b_0, b_1], & \text{when } q = 1.
\end{cases} \tag{5.4}$$

Here we have $b_0 = M_{\min}$, $b_Q = M_{\max}$. The break points $b_q$ are determined by dividing the interval of $[M_{\min}, M_{\max}]$ with equal distance, i.e.,

$$b_q = M_{\min} + q\Delta \tag{5.5}$$

where $\Delta$ is the bin length,

$$\Delta = \frac{M_{\max} - M_{\min}}{Q} \tag{5.6}$$

$$= b_q - b_{q-1}$$

$q = 1, 2, \ldots, Q$. In this study, I set $Q = 30$.

Step 2: Calculating BCQ.

BCQ is estimated as the proportion of correct pixels in each bin,

$$BCQ = \frac{\sum_{a=1}^{n} I(u_a)}{n} \tag{5.7}$$

where $I(u_a)$, as defined in Equation (5.3), is the indicator variable for pixel at location $u_a$, $n$ is the number of pixels in each bin. BCQ is actually the overall accuracy based on
pixels belonging to each bin. It can also be treated as the expected probability of a pixel being correctly classified. I have the following alternative definition for PPCC,

\[
\text{PPCC} \equiv \lim_{\Delta \to 0} \text{BCQ}
\]  

(5.8)

In practice, the difference of the lower \(b_{q-1}\) and upper \(b_q\) bound, i.e., \(\Delta = b_q - b_{q-1}\) for each bin is very small. It can be assumed that pixels in each bin have same MV equal to \(b_q\), and same classification quality as BCQ\(_q\).

### 2.3.2 Estimate PPCC

**Step 1:** Modeling relationship between BCQ\(_q\) vs. \(b_q\).

Based on the results from Chapter 4, there is strong empirical relationship between BCQ\(_q\) and \(b_q\). A mathematical model is fitted to the scatter plot of BCQ\(_q\) vs. \(b_q\) as:

\[
\text{BCQ}_q = f(b_q),
\]  

(5.9)

In practice, the fitted model \(f\) usually takes the form of a simple linear model.

**Step 2:** Estimate classification confidence.

Using the parameters estimated in equation (5.9), I predict the classification confidence for each pixel using the following piecewise function:

\[
\text{PPCC} = \begin{cases} 
1, & \text{if } f(\text{MV}) > 1 \\
0, & \text{if } f(\text{MV}) < 0 \\
f(\text{MV}), & \text{otherwise}
\end{cases}
\]  

(5.10)

Function \(f\) is not necessarily a bounded model. Therefore, it is possible that there are some extreme pixels whose \(f(\text{MV})\) exceed the range of \([0,1]\). If there do exist extreme pixels, I make an adjustment by resetting their PPCC values to 0 or 1. That is why I use a piecewise function in Equation (5.10). The adjustment does not have significant influence
on the estimates of PPCC. First, in practice, $f(MV)$ usually only exceed the range $[0,1]$ for a very small value. Second, the number of these extreme pixels is generally small.

2.4 Method evaluation

In Section 2.3, I proposed a regression method to estimate PPCC. In the following, I evaluate my method in four aspects: (1) Exploratory data analysis (EDA) using bi-histogram of PPCC for correct pixels and error pixels. (2) Statistical analysis of the relationship between binned per-pixel classification confidence (BPC) and PPCC. (3) Stability of PPCC estimates under simulated test data; (4) Impact of sample design on PPCC estimates.

2.4.1 Relationship between binned PPCC and PPCC

Ideally, the estimated PPCC should be evaluated with the reference PPCC. However, there is no such reference PPCC in practice. Therefore, I propose an alternative method to evaluate PPCC. I define a new statistic called binned per-pixel classification confidence (BPC) and then examine the relationship between BPC and PPCC. The procedures to calculate BPC is the same as that I calculate BCQ in Section 2.3.1. First, I group pixels into bins based on PPCC. Second, BPC is defined as the proportion of pixels correctly classified in each bin.

By definition, BPC for each bin represent the expected probability of each pixel in the bin being correctly classified. Recall that PPCC is defined as the probability of a pixel being correctly classified. Therefore, the following assumption holds,

$$PPCC \rightarrow \lim_{\Delta \rightarrow 0} BPC.$$  \hspace{1cm} (5.11)

In order to evaluation PPCC estimates, I draw a scatter plot of BPC against PPCC. In practice, when the bin length, $\Delta$, is small enough, I expect that BPC and its
corresponding PPCC should lie well along with the 1:1 reference line. I calculate the coefficient of determination, i.e., $R^2$, between BPC and PPCC to examine the agreement between BPC and PPCC.

2.4.2 Stability of PPCC estimates

The model estimated in Equation (5.9) above is based on $S_2$ test pixels. I propose a simulation method to examine the stability of PPCC estimates. Recall that the whole image I classified is composed of $S$ pixels and I have full coverage reference data for all the $S$ pixels. For the $S$ reference pixels, $S_1$ pixels have been used for training and $S_2$ pixels for testing. Using the remaining $S - S_1 - S_2$ pixels, I simulate another $T$ test datasets and run the method $T$ times to estimate PPCC accordingly. The variation of the estimates is then used to evaluate the performance of method. The detailed steps are shown below.

Step 1: Simulate test datasets.

I simulate $T$ test datasets using the remaining $S - S_1 - S_2$ pixels. Each of the $T$ test datasets is composed of $S_2$ pixels randomly drawn from $S - S_1 - S_2$ pixels without replacement.

Step 2: Calculate MV and BCQ.

For each of the $T$ test datasets, I calculate MV and BCQ. MV is extracted from the probability vectors for each test dataset using equation (5.2). BCQ is calculated using the same method as Step 1 and 2 in Section 2.3.

Step 3: Fit the model using simulated data.

For each of the $T$ test datasets, I get two sequences of data: $BCQ_t$ and $b_t$. Plot the scatter plot of $BCQ_t$ vs. $b_t$ and fit it with a mathematical model $BCQ_t = f_t(b_t)$, where $t = 1, 2, ..., T$. In total I get $T$ models.
Step 4: Estimate PPCC.

Estimate PPCC using the following models:

\[
PPCC_t = \begin{cases} 
1, & \text{if } f'(MV) > 1 \\
0, & \text{if } f'(MV) < 0 \\
f_t(MV), & \text{otherwise}
\end{cases}
\]

where \( t = 1, 2, \ldots, T \). In total I get \( T \) maps of PPCC.

Step 5: Calculate RMSE and Willmott's \( d \).

Calculate RMSE and Willmott's \( d \) using \( PPCC_t \) from equation (5.12) against the PPCC estimated in Section 2.3. RMSE and Willmott's \( d \) are two commonly used statistics to measure the agreement of two maps. RMSE is calculated using the following formula,

\[
RMSE_t = \left[ \frac{1}{S} \sum_{i=1}^{S} (PPCC(i) - PPCC_t(i))^2 \right]^{1/2}
\]

where \( PPCC \) is the estimated per-pixel classification confidence based on \( S_2 \) test pixel, \( PPCC_t \) the per-pixel classification confidence estimated using the \( t \)th simulated test dataset, and \( S \) is the total number of pixels in the image.

Willmott's \( d \) is a measure of agreement proposed by Willmott (Willmott 1981, 1982; Willmott et al. 2012), which is calculated as,

\[
d_t = 1 - \frac{\sum_{i=1}^{S} \left[ (PPCC(i) - PPCC_t(i))^2 \right]}{\sum_{i=1}^{S} \left[ (PPCC(i) - PPCC) - (PPCC_t(i) - PPCC) \right]^2}
\]

where \( PPCC = \frac{1}{S} \sum_{i=1}^{S} PPCC(i) \). Compared to RMSE, Willmott's \( d \) has two main advantages: (1) Bounded in a range from 0 to 1, indicating degree of agreement ranging
from complete disagreement to complete agreement. (2) Non-dimensional, thus easier to interpret than the widely used RMSE.

Step 6: Examine the relationship between BPC$_t$ and PPCC$_t$.

For each set of PPCC$_t$, calculate binned PPCC, i.e. BPC$_t$. Similar to Section 2.4.3, examine the relationship between BPC$_t$ and PPCC$_t$ and estimate the $R^2$.

2.5 Impact of sample design on PPCC estimates

2.5.1 Selection of sample design

Previous studies have examined the impact of sampling design on global accuracy assessment (Chen and Wei 2009; Congalton 1988a; Stehman and Foody 2009). In this study, I will examine the impact of sampling design on the estimation of local classification quality, i.e., PPCC. Sampling design includes three close related parts: sample unit, sample size, and sampling scheme.

Sample unit is the unit to select samples from reference map (Congalton and Green 2009). There are two most commonly used sample units: pixel and block of pixels (Broich et al. 2009; Stehman and Czaplewski 1998). The use of pixel unit is straightforward and causes no confusion. However, the use of block unit needs more explanation here. When a block of pixels is used as sample unit, the block can be of arbitrary shape. Without loss of generality, I use the square block of $3 \times 3$ pixels in this study.

Sample size is generally determined based on certain statistical principles, for example, the normal approximation of binomial distribution (Congalton and Green 2009; Foody 2009). In practice, a thumb of rule is commonly used, where a minimum of 50 samples are required for each class (Congalton and Green 2009; Ginevan 1979; Hay...
For our three datasets, the numbers of classes are 5-7. In this study, I examine the impact of total sample size on the estimate of PPCC under six levels: 630, 1260, 1890, 2520, 3150, 3780 pixels. The sample sizes are determined so that exact same number of samples can be obtained under both pixel sampling and cluster sampling.

Sampling scheme is statistically defined as a function to assign selection probabilities to sample units (de Gruijter et al. 2006). In other words, sampling schemes determines how sample are selected. In this study, I examine four commonly used basic sampling schemes (Congalton and Green 2009; Stehman 2009): (1) simple random sampling (SRS), (2) stratified random sampling (STS), (3) systematic sampling (SYS), and (4) cluster random sampling (CRS). SRS draws samples at random and mutually independent. STS first divide map into strata, and then conduct SRS in each stratum. In practice, there are two types of strata: class strata, spatial strata (Congalton 1988a; Stehman 2009). In this study, I use class a stratum, i.e., the map is divided into groups by classes. SYS draws samples from locations that are with fixed equal distance apart vertically and horizontally. The fixed equal distance varies with the sample sizes. CRS first divide the map into clusters and then randomly select a number of clusters as the sample pool. In this study, the clusters are formed by dividing the map into $3 \times 3$ square blocks. I use one stage CRS, where all the pixels in each pooled cluster are selected as samples.

In summary, the sample design on the estimate of PPCC is examined in the following aspects:

(1) Sampling scheme: SRS, STS, SYS, CRS.

(2) Sample size: 630, 1260, 1890, 2520, 3150, 3780 pixels.
(3) Sample unit: pixel unit for SRS, STS, and SYS; 3x3 pixel block for CRS.

The above experiment is applied to three datasets with four classifiers. In total, there are $3 \times 4 \times 4 \times 6 = 288$ scenarios.

2.5.2 Impact of sample design on PPCC

In this study, I examine the impact of sampling scheme and sample size on the estimate of PPCC. I have four sample schemes and six levels of sample size, resulting 24 scenarios. Under each scenario, I compare PPCC in two aspects:

(1) Relationship between BCQ and MV;

(2) Relationship between BPC and PPCC.

In the first aspect, BCQ and MV are used to fit a model which is then used to estimate PPCC. In the second aspect, BPC and PPCC are used to evaluate the estimated PPCC. In both aspects, the goodness of fit, $R^2$ is compared. I use ANOVA to reveal the influencing factors in the estimate of PPCC.

3 Results

3.1 Maps of classification error and MV

Figure 5.2 shows the indicator maps of classification error (by MLC) which are generated by comparing the class map with the full coverage reference map. For Data 1, clusters of classification error indicate that the main confusion is from two classes: young crop and bare soil. For Data 2 and 3, there exist confusion between concrete surface and roof. Figure 5.3 shows the spatial distribution of MV for three datasets based on MLC. There exist some spatial patterns for all three datasets. The clusters of low values in Figure 5.3 are consistent with the clusters of classification error in Figure 5.2.
Figure 5.2 Indicator maps of classification error (MLC)

Figure 5.3 Maps of MV based on MLC
3.2 Estimated PPCC

3.2.1 Estimated model

Figure 3.14 shows the scatter plots of BCQ$_q$ vs. $b_q$, where BCQ$_q$ are calculated using $S_2$ test data. A reference line of simple linear relationship is also drawn on each panel. Although the scatter plots are much dispersed than those based on full coverage reference data (Figure 3.21 in Chapter 3), there is strong positive relationship between BCQ$_q$ vs. $b_q$.

Figure 3.14 shows that a linear relationship may exist between BCQ$_q$ vs. $b_q$. Fit the scatter plots in Figure 3.14 with the following simple linear models:

$$BCQ_q = f(b_q) = a \times b_q + c$$  \hspace{1cm} (5.15)

and the model parameters are shown in Table 5.1. The model to estimate PPCC is now written as

$$PPCC = \begin{cases} 
1, & \text{if} \ a \times MV + c > 1 \\
0, & \text{if} \ a \times MV + c < 0 \\
a \times MV + c, & \text{otherwise}
\end{cases}$$  \hspace{1cm} (5.16)

Using equation (5.16) and the parameters $(a, c)$ in Table 5.1, I estimate PPCC for each pixel.
Table 5.1 Model parameters for BCQ$_q$ vs. $b_q$ based MV of test data

<table>
<thead>
<tr>
<th></th>
<th>MLC</th>
<th></th>
<th></th>
<th>NN</th>
<th></th>
<th></th>
<th>SVM</th>
<th></th>
<th></th>
<th>BDT</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a$</td>
<td>$c$</td>
<td>$R^2$</td>
<td>$a$</td>
<td>$c$</td>
<td>$R^2$</td>
<td>$a$</td>
<td>$c$</td>
<td>$R^2$</td>
<td>$a$</td>
<td>$c$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>Data 1</td>
<td>0.60</td>
<td>0.39</td>
<td>0.79</td>
<td>0.49</td>
<td>0.47</td>
<td>0.73</td>
<td>0.54</td>
<td>0.40</td>
<td>0.63</td>
<td>0.60</td>
<td>0.33</td>
<td>0.79</td>
</tr>
<tr>
<td>Data 2</td>
<td>0.55</td>
<td>0.45</td>
<td>0.82</td>
<td>0.56</td>
<td>0.45</td>
<td>0.85</td>
<td>0.63</td>
<td>0.34</td>
<td>0.77</td>
<td>0.58</td>
<td>0.40</td>
<td>0.86</td>
</tr>
<tr>
<td>Data 3</td>
<td>0.52</td>
<td>0.41</td>
<td>0.88</td>
<td>0.60</td>
<td>0.40</td>
<td>0.86</td>
<td>0.60</td>
<td>0.45</td>
<td>0.91</td>
<td>0.50</td>
<td>0.43</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Figure 5.4 Scatter plot of BCQ$_q$ vs. $b_q$ based on MV of test data
3.2.2 Characteristics of PPCC

The maps of estimated PPCC based on MLC are shown in Figure 5.5. There are clear spatial patterns for all three datasets. Visually compare the class map with the map of PPCC, I can identify the properties of PPCC clusters. For Data 1, there are some clusters of low PPCC for young crop. For Data 2, concrete surface, roof, and shadow generally have low PPCC. For Data 3, water has the highest PPCC, concrete surface and roof have the lowest PPCC.

Figure 5.6 - Figure 5.11 show the histogram of PPCC by class for three datasets which confirms the results from visual maps. For Data 1, the distribution of PPCC for all classes except young crop is negatively skewed. PPCC are mostly concentrated in high values. For young crop, PPCC are positively skewed. The pattern of PPCC values is consistent with the user's/producer's accuracies, where young crop has the lowest user's accuracy (0.51) and producer's accuracy (0.66). The overall accuracy for Data 1 with MLC is 0.88. For Data 2, PPCC for tree/shrub and bare soil are mostly high while PPCC for other classes are more spread. For Data 3, PPCC for tree/shrub, grass, and water are mostly high while PPCC for other classes are more spread. The results are consistent with the global accuracy measures, i.e., user's accuracy and producer's accuracy.
Figure 5.5 Predicted PPCC using MV based on MLC

(a) - Data 1
(b) - Data 2
(c) - Data 3

Figure 5.6 PPCC by class for Data 1 using MLC
Figure 5.7 PPCC by class for Data 1 using NN

Figure 5.8 PPCC by class for Data 1 using SVM
Figure 5.9 PPCC by class for Data 1 using BDT

Figure 5.10 PPCC by class for Data 2 using MLC
The results of PPCC for different classifiers are different. For example, for Data 1, the PPCC for young crop is positively skewed, while those for other classifiers are slightly negatively skewed. The different pattern of PPCC for different classifiers is all consistent with the global accuracy measures. Therefore, PPCC do catch the classification quality.

3.3 Evaluation of PPCC

3.3.1 Bi-histogram of PPCC for correct pixels and error pixels

Figure 5.12 shows the bi-histogram of PPCC for correct/error pixels. The distribution of correct pixels and error pixels are very different. PPCC of correct pixels are generally of high PPCC, while the PPCC of error pixels are more spread out. It also shows that the minimum of PPCC is high above zero. PPCC are mostly greater than 0.4 except for Data 1 with BDT and Data 2 with SVM and BDT, where PPCC are greater than 0.3.
3.3.2 Relationship between binned PPCC and PPCC

Figure 5.13 shows the scatter plot of binned PPCC (BPC) against PPCC. It is clear that BPC and PPCC align well on the 1:1 reference line. The $R^2$ values are mostly as high as above 0.95 except for Data 1 with SVM where $R^2 = 0.9041$. This confirms that PPCC is a good statistic for measuring the probability of pixels being correctly classified. In other words, PPCC does measure the per-pixel classification confidence.
3.3.3 Stability of PPCC estimates

Using simulated test datasets, I estimated maps of PPCC, based on \( t \)th test dataset. Table 5.2 shows the mean and standard deviation of RMSE, Willmott's \( d \), and \( R^2 \) indicating the stability of PPCC estimates in 100 iterations. RMSE and Willmott's \( d \) are calculated using simulated PPCC\(_t\) and estimated PPCC. RMSE are all very low which indicates that PPCC\(_t\) agree well with PPCC. Willmott's \( d \) are mostly high which indicates high agreement between PPCC\(_t\) and PPCC.

The relationship between BPC\(_t\) and PPCC\(_t\) align well with the 1:1 reference line. The mean and standard deviation of \( R^2 \) are mostly above 0.95, except for Data 1 with
SVM and BDT. The high values of $R^2$ confirm that BPC and PPCC fit well with the 1:1 reference line.

Table 5.2 Mean and standard deviation for RMSE, Willmott's $d$, and $R^2$

<table>
<thead>
<tr>
<th></th>
<th>MLC</th>
<th></th>
<th>NN</th>
<th></th>
<th>SVM</th>
<th></th>
<th>BDT</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>RMSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data 1</td>
<td>0.0194</td>
<td>0.0076</td>
<td>0.0221</td>
<td>0.0078</td>
<td>0.0174</td>
<td>0.0101</td>
<td>0.0201</td>
<td>0.0095</td>
</tr>
<tr>
<td>Data 2</td>
<td>0.0175</td>
<td>0.0090</td>
<td>0.0247</td>
<td>0.0085</td>
<td>0.0131</td>
<td>0.0065</td>
<td>0.0230</td>
<td>0.0094</td>
</tr>
<tr>
<td>Data 3</td>
<td>0.0184</td>
<td>0.0077</td>
<td>0.0246</td>
<td>0.0123</td>
<td>0.0310</td>
<td>0.0142</td>
<td>0.0279</td>
<td>0.0159</td>
</tr>
<tr>
<td>$d$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data 1</td>
<td>0.9957</td>
<td>0.0032</td>
<td>0.9943</td>
<td>0.0037</td>
<td>0.9942</td>
<td>0.0075</td>
<td>0.9954</td>
<td>0.0042</td>
</tr>
<tr>
<td>Data 2</td>
<td>0.9955</td>
<td>0.0047</td>
<td>0.9937</td>
<td>0.0039</td>
<td>0.9981</td>
<td>0.0019</td>
<td>0.9929</td>
<td>0.0054</td>
</tr>
<tr>
<td>Data 3</td>
<td>0.9964</td>
<td>0.0032</td>
<td>0.9909</td>
<td>0.0081</td>
<td>0.9812</td>
<td>0.0177</td>
<td>0.9862</td>
<td>0.0181</td>
</tr>
<tr>
<td>$R^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data 1</td>
<td>0.9834</td>
<td>0.0144</td>
<td>0.9763</td>
<td>0.0334</td>
<td>0.9144</td>
<td>0.0387</td>
<td>0.9404</td>
<td>0.0375</td>
</tr>
<tr>
<td>Data 2</td>
<td>0.9676</td>
<td>0.0145</td>
<td>0.9752</td>
<td>0.0120</td>
<td>0.9600</td>
<td>0.0193</td>
<td>0.9533</td>
<td>0.0195</td>
</tr>
<tr>
<td>Data 3</td>
<td>0.9695</td>
<td>0.0133</td>
<td>0.9842</td>
<td>0.0158</td>
<td>0.9760</td>
<td>0.0114</td>
<td>0.9684</td>
<td>0.0220</td>
</tr>
</tbody>
</table>

3.4 Impact of sampling design on PPCC estimates

3.4.1 Relationship between BCQ and MV

Figure 5.14 - Figure 5.17 show the scatter plots of BCQ$_q$ vs. $b_q$ with MLC and 630 test pixels under different sampling schemes. The difference between Figure 5.14 and Figure 3.14 reveals the impact of sample size on PPCC estimate. Both Figure 3.14 and Figure 5.14 are generated based on SRS, but with 2000 and 630 test pixels respectively. With less test pixels, points on the scatter plot of BCQ$_q$ vs. $b_q$ are more scattered. The
performances of different classifiers are different. BDT are especially affected by sample size, while MLC, NN, and SVM are less affected by sample size.

Figure 5.14 Scatter plot of BCQ\(_q\) vs. \(b_q\) with MLC for 630 test pixels: SRS
Figure 5.15 Scatter plot of $\text{BCQ}_q$ vs. $b_q$ with MLC for 630 test pixels: STS

Figure 5.16 Scatter plot of $\text{BCQ}_q$ vs. $b_q$ with MLC for 630 test pixels: SYS
Table 5.3 shows the $R^2$ for fitting $BCQ_q$ vs. $b_q$ with simple linear model and sample size = 630. The low $R^2$ values confirm the visual results from the scatter plots in Figure 5.14 - Figure 5.17. As the points in the scatter plots for BDT are much scattered, the $R^2$ values for BDT are very low. In other words, the relationship between $BCQ_q$ and $b_q$ is not clear. Especially for Data 1 and 3 with BDT, the scatter plots look random and there is no relationship between $BCQ_q$ and $b_q$. Therefore, we get $R^2 = 0$. The impacts of sampling schemes vary across datasets and classifiers. There is no dominant sampling scheme that performs consistently better.

I examined the $R^2$ for fitting $BCQ_q$ vs. $b_q$ with simple linear model under different sample sizes. I found that when sample size = 1260, the results are much better than those
with sample size = 630. However, there are still some $R^2$ lower than 0.5. When sample size is no less than 1890, $R^2$ are all above 0.5 (Table 5.4).

Table 5.3 $R^2$ for fitting BCQ$_q$ vs. $b_q$ with simple linear model: sample size = 630

<table>
<thead>
<tr>
<th></th>
<th>Data 1</th>
<th></th>
<th></th>
<th>Data 2</th>
<th></th>
<th></th>
<th>Data 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRS</td>
<td>STS</td>
<td>SYS</td>
<td>CRS</td>
<td>SRS</td>
<td>STS</td>
<td>SYS</td>
<td>CRS</td>
</tr>
<tr>
<td>MLC</td>
<td>0.670</td>
<td>0.553</td>
<td>0.727</td>
<td>0.740</td>
<td>0.605</td>
<td>0.600</td>
<td>0.665</td>
<td>0.668</td>
</tr>
<tr>
<td>NN</td>
<td>0.634</td>
<td>0.334</td>
<td>0.378</td>
<td>0.723</td>
<td>0.641</td>
<td>0.551</td>
<td>0.700</td>
<td>0.568</td>
</tr>
<tr>
<td>SVM</td>
<td>0.532</td>
<td>0.601</td>
<td>0.490</td>
<td>0.767</td>
<td>0.307</td>
<td>0.719</td>
<td>0.700</td>
<td>0.650</td>
</tr>
<tr>
<td>BDT</td>
<td>0.063</td>
<td>0.088</td>
<td>0.121</td>
<td>0.000</td>
<td>0.501</td>
<td>0.198</td>
<td>0.151</td>
<td>0.160</td>
</tr>
</tbody>
</table>

Table 5.4 $R^2$ for fitting BCQ$_q$ vs. $b_q$ with simple linear model: sample size = 1890

<table>
<thead>
<tr>
<th></th>
<th>Data 1</th>
<th></th>
<th></th>
<th>Data 2</th>
<th></th>
<th></th>
<th>Data 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRS</td>
<td>STS</td>
<td>SYS</td>
<td>CRS</td>
<td>SRS</td>
<td>STS</td>
<td>SYS</td>
<td>CRS</td>
</tr>
<tr>
<td>MLC</td>
<td>0.824</td>
<td>0.858</td>
<td>0.731</td>
<td>0.727</td>
<td>0.855</td>
<td>0.828</td>
<td>0.722</td>
<td>0.758</td>
</tr>
<tr>
<td>NN</td>
<td>0.743</td>
<td>0.625</td>
<td>0.860</td>
<td>0.647</td>
<td>0.907</td>
<td>0.850</td>
<td>0.834</td>
<td>0.909</td>
</tr>
<tr>
<td>SVM</td>
<td>0.850</td>
<td>0.658</td>
<td>0.608</td>
<td>0.740</td>
<td>0.842</td>
<td>0.604</td>
<td>0.809</td>
<td>0.632</td>
</tr>
<tr>
<td>BDT</td>
<td>0.715</td>
<td>0.733</td>
<td>0.583</td>
<td>0.608</td>
<td>0.866</td>
<td>0.682</td>
<td>0.894</td>
<td>0.686</td>
</tr>
</tbody>
</table>

Figure 5.18 shows the impact of sample design on the relationship between BCQ$_q$ and $b_q$. The fitted $R^2$ generally increase with the sample size. When sample size is small, the difference of $R^2$ for different sampling schemes is large. When sample size is large,
the impact of sampling scheme can be ignored except for Data 2 with SVM. The performance of BDT depends more on sample size.

![Graph showing impact of sample design on the relationship between BCQ_q and b_q](image)

**Figure 5.18** Impact of sample design on the relationship between BCQ_q and b_q

Note: Sample size level 1-6 corresponding to sample size 630, 1260, 1890, 2520, 3150, 3780 pixels respectively.

Table 5.5 shows two-way ANOVA results which indicate the impact of sampling scheme and sample size on the relationship between BCQ_q vs. b_q, and thus the estimate of PPCC. For all three datasets and all four classifiers, sample size is a significant factor on
$R^2$ while sampling scheme is not significant. No confounding effects of sampling scheme and sample size are found except for Data 1 with NN and SVM. In addition, the large contrast of means of sum of squares (MSS) between sample size and sampling scheme shows that sample size does account for the estimate of PPCC.

Table 5.6 shows three-way ANOVA results which compare the impact of sampling scheme, sample size, and classifiers on the estimate of PPCC. For all three datasets, sample size, classifier, and their interaction are significant factors on the estimate of PPCC while sampling scheme is not significant. No confounding effects of sampling scheme with other two factors are found except for Data 1 with MLC. In addition, the large contrast of MSS between sample size and classifier confirms that sample size is the main factors influencing the estimate of PPCC.
### Table 5.5 Two-way ANOVA results for $R^2$ of BCQ\(q\) vs. $b_q$

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Factors</th>
<th>Df</th>
<th>Data 1 MSS</th>
<th>P-value</th>
<th>Data 2 MSS</th>
<th>P-value</th>
<th>Data 3 MSS</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLC</td>
<td>$X_1$</td>
<td>3</td>
<td>0.0016</td>
<td>0.6440</td>
<td>0.0018</td>
<td>0.7590</td>
<td>0.0095</td>
<td>0.5620</td>
</tr>
<tr>
<td></td>
<td>$X_2$</td>
<td>1</td>
<td>0.1228</td>
<td>0.0000</td>
<td>0.1470</td>
<td>0.0000</td>
<td>0.3650</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>$X_1^*X_2$</td>
<td>3</td>
<td>0.0037</td>
<td>0.2960</td>
<td>0.0012</td>
<td>0.8520</td>
<td>0.0157</td>
<td>0.3540</td>
</tr>
<tr>
<td>Residuals</td>
<td></td>
<td>16</td>
<td>0.0027</td>
<td>0.0000</td>
<td>0.0044</td>
<td></td>
<td>0.0135</td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td>$X_1$</td>
<td>3</td>
<td>0.0115</td>
<td>0.1647</td>
<td>0.0018</td>
<td>0.8330</td>
<td>0.0050</td>
<td>0.5830</td>
</tr>
<tr>
<td></td>
<td>$X_2$</td>
<td>1</td>
<td>0.2814</td>
<td>0.0000</td>
<td>0.2175</td>
<td>0.0000</td>
<td>0.3924</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>$X_1^*X_2$</td>
<td>3</td>
<td>0.0290</td>
<td>0.0137</td>
<td>0.0018</td>
<td>0.8370</td>
<td>0.0025</td>
<td>0.8040</td>
</tr>
<tr>
<td>Residuals</td>
<td></td>
<td>16</td>
<td>0.0060</td>
<td></td>
<td>0.0062</td>
<td></td>
<td>0.0075</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>$X_1$</td>
<td>3</td>
<td>0.0075</td>
<td>0.3587</td>
<td>0.0183</td>
<td>0.3108</td>
<td>0.0010</td>
<td>0.8860</td>
</tr>
<tr>
<td></td>
<td>$X_2$</td>
<td>1</td>
<td>0.2279</td>
<td>0.0000</td>
<td>0.1243</td>
<td>0.0092</td>
<td>0.1436</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>$X_1^*X_2$</td>
<td>3</td>
<td>0.0194</td>
<td>0.0619</td>
<td>0.0173</td>
<td>0.3346</td>
<td>0.0041</td>
<td>0.4760</td>
</tr>
<tr>
<td>Residuals</td>
<td></td>
<td>16</td>
<td>0.0065</td>
<td></td>
<td>0.0142</td>
<td></td>
<td>0.0047</td>
<td></td>
</tr>
<tr>
<td>BDT</td>
<td>$X_1$</td>
<td>3</td>
<td>0.0171</td>
<td>0.5810</td>
<td>0.0110</td>
<td>0.7460</td>
<td>0.0201</td>
<td>0.3180</td>
</tr>
<tr>
<td></td>
<td>$X_2$</td>
<td>1</td>
<td>1.4844</td>
<td>0.0000</td>
<td>0.9904</td>
<td>0.0000</td>
<td>1.4937</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>$X_1^*X_2$</td>
<td>3</td>
<td>0.0088</td>
<td>0.7920</td>
<td>0.0149</td>
<td>0.6510</td>
<td>0.0074</td>
<td>0.7100</td>
</tr>
<tr>
<td>Residuals</td>
<td></td>
<td>16</td>
<td>0.0254</td>
<td></td>
<td>0.0267</td>
<td></td>
<td>0.0158</td>
<td></td>
</tr>
</tbody>
</table>

$X_1$: sampling scheme, $X_2$: sample size. Sig. level: ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 .

### Table 5.6 Three-way ANOVA results for $R^2$ of BCQ\(q\) vs. $b_q$

<table>
<thead>
<tr>
<th>Factors</th>
<th>Df</th>
<th>Data 1 MSS</th>
<th>P-value</th>
<th>Data 2 MSS</th>
<th>P-value</th>
<th>Data 3 MSS</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>3</td>
<td>0.0053</td>
<td>0.6666</td>
<td>0.0104</td>
<td>0.4939</td>
<td>0.0158</td>
<td>0.2180</td>
</tr>
<tr>
<td>$X_2$</td>
<td>1</td>
<td>1.6598</td>
<td>0.0000</td>
<td>1.2072</td>
<td>0.0000</td>
<td>2.0046</td>
<td>0.0000</td>
</tr>
<tr>
<td>$X_3$</td>
<td>3</td>
<td>0.1712</td>
<td>0.0000</td>
<td>0.0540</td>
<td>0.0090</td>
<td>0.2202</td>
<td>0.0000</td>
</tr>
<tr>
<td>$X_1^*X_2$</td>
<td>3</td>
<td>0.0296</td>
<td>0.0409</td>
<td>0.0022</td>
<td>0.9154</td>
<td>0.0104</td>
<td>0.3980</td>
</tr>
<tr>
<td>$X_1^*X_3$</td>
<td>9</td>
<td>0.0108</td>
<td>0.4023</td>
<td>0.0075</td>
<td>0.8074</td>
<td>0.0066</td>
<td>0.7600</td>
</tr>
<tr>
<td>$X_2^*X_3$</td>
<td>3</td>
<td>0.1522</td>
<td>0.0000</td>
<td>0.0906</td>
<td>0.0004</td>
<td>0.1300</td>
<td>0.0000</td>
</tr>
<tr>
<td>$X_1^*X_2^*X_3$</td>
<td>9</td>
<td>0.0104</td>
<td>0.4287</td>
<td>0.0110</td>
<td>0.5728</td>
<td>0.0064</td>
<td>0.7760</td>
</tr>
<tr>
<td>Residuals</td>
<td>64</td>
<td>0.0101</td>
<td></td>
<td>0.0129</td>
<td></td>
<td>0.0104</td>
<td></td>
</tr>
</tbody>
</table>

$X_1$: sample scheme, $X_2$: sample size, $X_3$: classifier. Sig. level: ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
3.4.2 Relationship between BPC and PPCC

Figure 5.19 - Figure 5.22 show the scatter plots of BPC vs. PPCC with MLC and 630 test pixels under different sampling schemes. The difference between Figure 5.19 and Figure 5.13 reveals the impact of sample size on PPCC estimate. Both Figure 5.19 and Figure 5.13 are generated based on SRS, but with 2000 and 630 test pixels respectively. With less test pixels, scatter plots of BPC vs. PPCC no longer align with the 1:1 reference line. MLC is relatively robust in estimating PPCC except for Data 1 with SRS, and Data 3 with CRS. NN is relatively robust in estimating PPCC except for Data 3 and Data 2 with STS. The performance of SVM in estimating PPCC is more dependent on dataset and sampling scheme when sample size is small. BDT are most affected by sample size. When sample size = 630, we cannot obtain reliable PPCC from BDT.

![Scatter plots of BPC vs. PPCC with different models and datasets](image)

Figure 5.19 Scatter plot of BPC vs. PPCC with MLC for 630 test pixels: SRS
Figure 5.20 Scatter plot of BPC vs. PPCC with MLC for 630 test pixels: STS

Figure 5.21 Scatter plot of BPC vs. PPCC with MLC for 630 test pixels: SYS
It should be noted that in Figure 5.22, for Data 1 and 3 with BDT, the scatter plots of BPC vs. PPCC follows a vertical line. This is due to the effect shown in Figure 5.17 where there is no relationship between $B_{CQ_q}$ and $b_q$. This leads to meaningless PPCC estimation. Therefore, when sample size is too small, BDT is not suitable for estimating PPCC.

Table 5.7 shows the $R^2$ for fitting BPC vs. PPCC with the 1:1 reference line with sample size = 630. The variability in $R^2$ values confirms the visual results from the scatter plots in Figure 5.19 - Figure 5.22. As the points in the scatter plots for BDT are much deviated from the 1:1 reference line, the $R^2$ values for BDT are very low. In other words, the estimated PPCCs do not characterize classification quality at pixel level.
I examined the scatter plots of BPC vs. PPCC and the $R^2$ values for fitting BPC vs. PPCC with 1:1 reference line under different sample sizes. I found that when sample size = 1260, the results are much better than those with sample size = 630. However, there are still some cases where the scatter plots do not align well with the 1:1 reference line and $R^2$ lower than 0.5. When sample size is no less than 1890, the scatter plots align well with the 1:1 reference line and the $R^2$ values are all above 0.5 (Table 5.8).

Table 5.7 $R^2$ for fitting BPC vs. PPCC with the 1:1 reference line: sample size = 630

<table>
<thead>
<tr>
<th></th>
<th>Data 1</th>
<th></th>
<th>Data 2</th>
<th></th>
<th>Data 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRS</td>
<td>STS</td>
<td>SYS</td>
<td>CRS</td>
<td>SRS</td>
</tr>
<tr>
<td>MLC</td>
<td>0.885</td>
<td>0.986</td>
<td>0.994</td>
<td>0.998</td>
<td>0.961</td>
</tr>
<tr>
<td>NN</td>
<td>0.992</td>
<td>0.985</td>
<td>0.953</td>
<td>0.976</td>
<td>0.981</td>
</tr>
<tr>
<td>SMV</td>
<td>0.986</td>
<td>0.977</td>
<td>0.955</td>
<td>0.907</td>
<td>0.850</td>
</tr>
<tr>
<td>BDT</td>
<td>0.677</td>
<td>0.749</td>
<td>0.762</td>
<td>0.000</td>
<td>0.823</td>
</tr>
</tbody>
</table>

Table 5.8 $R^2$ for fitting BPC vs. PPCC with simple linear model: sample size = 1890

<table>
<thead>
<tr>
<th></th>
<th>Data 1</th>
<th></th>
<th>Data 2</th>
<th></th>
<th>Data 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRS</td>
<td>STS</td>
<td>SYS</td>
<td>CRS</td>
<td>SRS</td>
</tr>
<tr>
<td>MLC</td>
<td>0.993</td>
<td>0.994</td>
<td>0.981</td>
<td>0.973</td>
<td>0.970</td>
</tr>
<tr>
<td>NN</td>
<td>0.952</td>
<td>0.973</td>
<td>0.975</td>
<td>0.964</td>
<td>0.974</td>
</tr>
<tr>
<td>SMV</td>
<td>0.978</td>
<td>0.952</td>
<td>0.957</td>
<td>0.989</td>
<td>0.985</td>
</tr>
<tr>
<td>BDT</td>
<td>0.967</td>
<td>0.913</td>
<td>0.950</td>
<td>0.929</td>
<td>0.981</td>
</tr>
</tbody>
</table>
Figure 5.23 shows the impact of sample design on the relationship between BPC and PPCC. The impact of sample design differs across data and classifiers. The results for BDT are quite different from those for MLC, NN, and SVM. For BDT, $R^2$ values increase quickly with the increase of sample size and level off at sample level 3, i.e., sample size = 1890. The change of $R^2$ is less significant for other classifiers. The performance of CSR has larger variance than other sampling schemes for BDT.

Figure 5.23 Impact of sample design on the relationship between BPC vs. PPCC

Note: Sample size level 1-6 corresponding to sample size 630, 1260, 1890, 2520, 3150, 3780 pixels respectively.
In order to examine the performance of other classifiers, a zoom plot is shown in Figure 5.24. For Data 1 with MLC and SRS and Data 3 with MLC and SVM, $R^2$ values increase with sample size and level off after sample level 2, i.e., sample size = 1260.
There is no clear pattern for the variability of $R^2$ in other scenarios. Generally, $R^2$ has larger variance under CRS than other sampling schemes.

Table 5.9 shows two-way ANOVA results which indicate the impact of sampling scheme and sample size on the relationship between BPC vs. PPCC. The relationship between BPC and PPCC is robust in most scenarios except for the following scenarios: (1) For all three datasets with BDT and Data 3 with all classifiers, sample size is a significant factor influencing relationship between BPC vs. PPCC; (2) For Data 3 with NN, BDT, and Data 2 with MLC, sampling scheme is significant; (3) For Data 1 with BDT, Data 3 with NN, BDT, there exist confounding effects between sampling scheme and sample size. In addition, the MSS of all significant factors are also much higher that for non-significant factors.

Table 5.10 shows three-way ANOVA results which compare the impact of sampling scheme, sample size, and classifiers on the relationship between BPC vs. PPCC. For all three datasets, sample size, classifier, and their interaction are significant factors on the relationship between BPC vs. PPCC while sampling scheme is only significant for Data 2 and 3. The confounding effects of sampling scheme and sample size are significant with Data 1 and 3 while the confounding effects between sampling scheme and classifiers are not significant for all three datasets. The confounding effects of all three factors are significant for Data 1 AND Data 3. In addition, the large contrast of MSS between sample size and classifier confirms that sample size is the main factors influencing the relationship between PPCC.
Table 5.9 Two-way ANOVA results for $R^2$ of BPC vs. PPCC

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Factors</th>
<th>Df</th>
<th>Data 1 MSS</th>
<th>P-value</th>
<th>Data 2 MSS</th>
<th>P-value</th>
<th>Data 3 MSS</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLC</td>
<td>$X_1$</td>
<td>3</td>
<td>0.0006</td>
<td>0.3320</td>
<td>0.0020</td>
<td>0.0202 *</td>
<td>0.0044</td>
<td>0.3244</td>
</tr>
<tr>
<td></td>
<td>$X_2$</td>
<td>1</td>
<td>0.0014</td>
<td>0.1150</td>
<td>0.0001</td>
<td>0.6331</td>
<td>0.0202</td>
<td>0.0287 *</td>
</tr>
<tr>
<td></td>
<td>$X_1\times X_2$</td>
<td>3</td>
<td>0.0005</td>
<td>0.4470</td>
<td>0.0001</td>
<td>0.9525</td>
<td>0.0068</td>
<td>0.1631</td>
</tr>
<tr>
<td>Residuals</td>
<td></td>
<td>16</td>
<td>0.0005</td>
<td>0.0000</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0035</td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td>$X_1$</td>
<td>3</td>
<td>0.0001</td>
<td>0.7960</td>
<td>0.0022</td>
<td>0.1090</td>
<td>0.0060</td>
<td>0.0766 .</td>
</tr>
<tr>
<td></td>
<td>$X_2$</td>
<td>1</td>
<td>0.0004</td>
<td>0.2090</td>
<td>0.0017</td>
<td>0.1930</td>
<td>0.0405</td>
<td>0.0005 ***</td>
</tr>
<tr>
<td></td>
<td>$X_1\times X_2$</td>
<td>3</td>
<td>0.0002</td>
<td>0.4510</td>
<td>0.0007</td>
<td>0.5510</td>
<td>0.0086</td>
<td>0.0268 *</td>
</tr>
<tr>
<td>Residuals</td>
<td></td>
<td>16</td>
<td>0.0002</td>
<td></td>
<td>0.0009</td>
<td></td>
<td>0.0022</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>$X_1$</td>
<td>3</td>
<td>0.0008</td>
<td>0.3440</td>
<td>0.0003</td>
<td>0.8950</td>
<td>0.0005</td>
<td>0.8425</td>
</tr>
<tr>
<td></td>
<td>$X_2$</td>
<td>1</td>
<td>0.0009</td>
<td>0.2580</td>
<td>0.0009</td>
<td>0.4590</td>
<td>0.0336</td>
<td>0.0005 ***</td>
</tr>
<tr>
<td></td>
<td>$X_1\times X_2$</td>
<td>3</td>
<td>0.0004</td>
<td>0.6060</td>
<td>0.0034</td>
<td>0.1330</td>
<td>0.0008</td>
<td>0.7397</td>
</tr>
<tr>
<td>Residuals</td>
<td></td>
<td>16</td>
<td>0.0007</td>
<td></td>
<td>0.0016</td>
<td></td>
<td>0.0018</td>
<td></td>
</tr>
<tr>
<td>BDT</td>
<td>$X_1$</td>
<td>3</td>
<td>0.0363</td>
<td>0.2145</td>
<td>0.0157</td>
<td>0.3105</td>
<td>0.0595</td>
<td>0.0825 .</td>
</tr>
<tr>
<td></td>
<td>$X_2$</td>
<td>1</td>
<td>0.3403</td>
<td>0.0011 **</td>
<td>0.1732</td>
<td>0.0017 **</td>
<td>0.6505</td>
<td>0.0001 ***</td>
</tr>
<tr>
<td></td>
<td>$X_1\times X_2$</td>
<td>3</td>
<td>0.0610</td>
<td>0.0737 .</td>
<td>0.0110</td>
<td>0.4617</td>
<td>0.0707</td>
<td>0.0528 .</td>
</tr>
<tr>
<td>Residuals</td>
<td></td>
<td>16</td>
<td>0.0218</td>
<td></td>
<td>0.0122</td>
<td></td>
<td>0.0223</td>
<td></td>
</tr>
</tbody>
</table>

$X_1$: sampling scheme, $X_2$: sample size. Sig. level: ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1.

Table 5.10 Three-way ANOVA results for $R^2$ of BPC vs. PPCC

<table>
<thead>
<tr>
<th>Factors</th>
<th>Df</th>
<th>Data 1 MSS</th>
<th>P-value</th>
<th>Data 2 MSS</th>
<th>P-value</th>
<th>Data 3 MSS</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>3</td>
<td>0.0117</td>
<td>0.1211</td>
<td>0.0108</td>
<td>0.0437 *</td>
<td>0.0364</td>
<td>0.0040 **</td>
</tr>
<tr>
<td>$X_2$</td>
<td>1</td>
<td>0.1121</td>
<td>0.0001 ***</td>
<td>0.0620</td>
<td>0.0001 ***</td>
<td>0.4444</td>
<td>0.0000 ***</td>
</tr>
<tr>
<td>$X_3$</td>
<td>3</td>
<td>0.0812</td>
<td>0.0000 ***</td>
<td>0.0181</td>
<td>0.0045 **</td>
<td>0.0727</td>
<td>0.0000 ***</td>
</tr>
<tr>
<td>$X_1\times X_2$</td>
<td>3</td>
<td>0.0148</td>
<td>0.0630 .</td>
<td>0.0035</td>
<td>0.4396</td>
<td>0.0472</td>
<td>0.0008 ***</td>
</tr>
<tr>
<td>$X_1\times X_3$</td>
<td>9</td>
<td>0.0087</td>
<td>0.1666</td>
<td>0.0031</td>
<td>0.5901</td>
<td>0.0113</td>
<td>0.1595</td>
</tr>
<tr>
<td>$X_2\times X_3$</td>
<td>3</td>
<td>0.0770</td>
<td>0.0000 ***</td>
<td>0.0380</td>
<td>0.0000 ***</td>
<td>0.1001</td>
<td>0.0000 ***</td>
</tr>
<tr>
<td>$X_1\times X_2\times X_3$</td>
<td>9</td>
<td>0.0158</td>
<td>0.0095 **</td>
<td>0.0039</td>
<td>0.4290</td>
<td>0.0133</td>
<td>0.0892 .</td>
</tr>
<tr>
<td>Residuals</td>
<td>64</td>
<td>0.0058</td>
<td></td>
<td>0.0038</td>
<td></td>
<td>0.0074</td>
<td></td>
</tr>
</tbody>
</table>

$X_1$: sampling scheme, $X_2$: sample size, $X_3$: classifier. ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1.
3.4.3 Summary the effects of sample design

In summary, for my three datasets, sample size is the most important factor influencing the estimate of PPCC. Reliable estimate of PPCC can only be obtained with enough sample size, e.g., \( N \geq 1890 \) in our case.

The second factor influencing the estimate of PPCC is classifier. Due to the different algorithms using in different classifiers, the estimated PPCCs are a little different. However, as the large contrast of MSS for sample size and classifier shows, the influence of classifiers on the estimate of PPCC is relatively small.

The impact of sampling scheme on the estimate of PPCC can be ignored as it is either not significant or the MSS is very small compared to the MSS of sample size and classifier.

4 Discussion and Conclusions

In this chapter, I proposed a new method to estimate spatial classification quality using classification scores from classifiers. I defined two new statistics in this study: binned classification quality (BCQ) and per-pixel classification confidence (PPCC). BCQ measures the expected probability of a pixel being correctly classified and is calculated by grouping test pixels into bins based on the value of MV (margin of victory). I estimated a mathematical model between BCQ and MV based on test pixels and then applied the estimated model to all the pixels to predict PPCC. The proposed method is validated in the following four points.

First, bi-histogram of PPCC for correct pixels and error pixels. The purpose for this EDA is to examine if PPCC can distinguish error pixels from correct pixels. Based on the assumption, correct pixels should generally have high PPCC, while error pixels have low
PPCC. The bi-histogram showed that PPCCs for correct pixels are mainly of high values, while PPCCs for error pixels are more spread across the range. The results confirmed the assumptions about PPCC for correct pixels. The distribution of PPCC for error pixels is not clustered in low PPCC values, which may be due to the difficulty of classifiers in distinguishing classes. More exploration of the distribution of PPCC for error pixels is needed in the future.

Second, the relationship between binned PPCC (BPC) and PPCC. The purpose for this second evaluation is to test the performance of the method in the whole image. The mathematical model to predict PPCC is estimated based on BCQ and MV of test pixels. Ideally, I should test the predicted PPCC with real PPCC for the whole image. However, no real full coverage PPCC is available. The relationship between BPC and PPCC is an alternative approach to evaluate the method using full coverage reference data. The scatter plots of BPC vs. PPCC aligned well with the 1:1 reference line, which confirmed that PPCC did measure classification quality at pixel level.

Third, the stability of PPCC estimates using simulated test datasets. I estimated 100 versions of PPCC, using 100 test datasets. I then compared PPCC with the original PPCC by computing two statistics, RMSE and Willmott's $d$. The results showed that they agree with each other very well, which indicated that the estimation of PPCC is stable.

Fourth, the impact of sample design on PPCC estimate. I examine the robustness of PPCC estimate under four sampling schemes and six levels of sample size. The results showed that sample size is the most important factor influencing PPCC. The estimate of PPCC is robust when sample size is above certain level. When sample size is too small, no reliable PPCC estimates are available. The impact of sampling schemes on PPCC
estimates is negligible. The estimated PPCCs are comparable across classifiers and thus indicate the performance of different classifiers.

The essence of my method is to convert MV into PPCC through a regression model derived from test data. In all my test cases, a simple linear model was used. In practice, the readers can use any models that best fit their data. It should be emphasized that the transformation from MV to PPCC is not trivial. Margin of victory (MV), along with other probability measures such as primate probability (PP), relative entropy (RH), only carries the information of the potential capacity of a classifier to distinguish the spectral space into classes. The information of potential distinguishing capacity from these probability measures depends heavily on the training data and classification algorithm. We cannot directly use MV, PP, or RH as measures for classification quality for two reasons. First, these probability measures are generated using training data, but they are not validated with test data. We cannot tell to what extent they approach the truth. Second, they are not comparable across classifiers. This second limitation is closely related to the first one. For traditional accuracy assessment, the quantities for those global accuracy indices such as overall accuracy, producer's accuracy, user's accuracy, and kappa are comparable across different classifiers because they are estimated using test data. However, the quantities of probability measures are not comparable across classifiers due to the difference in the classification algorithms. The PPCC estimated using my method provide a vehicle to compare spatial classification quality from different classifiers.

The range of PPCC values needs some discussion here. As I mentioned in Section 2.3.2, the model $f$ converting MV to PPCC may not be a bounded model. The estimated $f(MV)$ may not be confined in the range of $[0,1]$ which is evident for the linear model
used in this study. In practice, there are only quite a small number of pixels whose estimated \( f(MV) \) exceed the bound of \([0, 1]\). It is safe to adjust the PPCC of these extreme pixels to 0 or 1. Much focus should be directed to the lower bound of \( f(MV) \).

As demonstrated in this study, the minimum of estimated PPCC are mainly around 0.4, with the lowest PPCC = 0.3283. This means that even for the worst distinguishable pixel, the probability of its being correctly classified is no less than 0.3283, far beyond zero. The high lower bound of PPCC beyond zero is valid and is the truth we have to accept. This phenomenon is due to the harshness in image classification that even the worst distinguishable pixels can have high margin of victory, primate probability, or relative entropy. In other words, we cannot classify an image with one hundred percent confidence. This also confirms that MV, PP, RH are not good measures for spatial classification quality, while PPCC estimated using my method is.

I tested and validated my new method on three representative datasets with four commonly used classifiers. The full coverage reference map made the evaluation procedures possible. Ideally, I should test the proposed method using as many images as possible, especially using images with various characteristics. However, it is not practical to do so using as many as possible real remote sensing images because a full coverage reference map is needed for each test image. A possible approach is to use simulated remote sensing images (Burnicki 2012; Chen et al. 2010; Liu and Chun 2009). Although simulating spectral images are difficult (Liu and Chun 2009), it is worth trying. In the future studies, I plan to conduct a comprehensive evaluation of my method using simulated spectral images.
CHAPTER 6 CONCLUSIONS AND IMPLICATIONS

1 Conclusions

Per-pixel classification confidence has been an important topic in both geographical information science (GIS) and remote sensing (RS) for three decades (Campbell 1981; Comber et al. 2012; Congalton 1988b; Foody 2002; Goodchild and Gopal 1989; Veregin 1996). Spatial variation of classification confidence will propagate to applications based on error-infected maps. For example, in the case of land cover change study, the spatial map quality for each input map will propagate to the land cover change map. Identify spatial classification confidence is thus play an important role in spatial data analysis. In this dissertation, I first evaluated previous methods on per-pixel classification confidence and then developed a regression model using classification scores from classifiers to estimate per-pixel classification confidence. Specifically, the dissertation includes four main chapters: literature review, evaluating classification score based method, evaluating interpolation based method, and estimating per-pixel classification confidence.

The literature review of previous method paved the ground for my study. In Chapter 2, I provided a comprehensive literature review on previous studies, including source of classification error and methods to estimate classification confidence. Previous scholars have studied the source of classification error both in single-date classified map and land cover change map. Methods used to identify classification error include visual
comparison and statistical regression. Statistical models include log-linear model, linear mixed model, and logistic regression model. Factors impacting spatial distribution of classification error can be summarized into three groups:

(1) System factors: sensor, platform, radiometric and geometric rectification;
(2) Object factors: topography, land use land cover pattern, landscape complexity;
(3) Interpretation factors: classification algorithm, sample design.

After I reviewed the literature on error source, I went forward to classify previous methods on classification confidence into three categories: classification score based method (CSBM), interpolation based method (IBM), and regression based method (RBM). Classification score based method derives classification confidence at pixel level using classification scores from classifiers. Interpolation based method generates per-pixel classification confidence by interpolating estimates at sample locations to the whole image. Regression based method identifies factors impacting classification error and then fits a statistical regression model of these factors to predict classification confidence. At the end of Chapter 2, I gave a brief summary of the advantages and limitations of previous studies, which are then examined in more detail in chapters 3 and 4: the evaluation of previous methods. The literature review concluded that previous studies had not been rigorously validated with empirical evidence. This gap calls for an comprehensive evaluation of previous methods.

Evaluation of previous methods not only filled the existing gap, but also laid foundation for my study to propose the new method. As discussed in Chapter 2, the regression based method has two limitations: (1) Data for predictors may not be easily available; (2) Model selection may be difficult. Due to the lack of data, this dissertation
did not evaluate regression based method. Therefore, the evaluation part of my dissertation focused on classification score based method (Chapter 3) and interpolation based method (Chapter 4). My evaluation concluded that interpolation based methods produce poor results due to interpolation effect. The classification score based method is promising.

Based on the suggestions from Chapter 3 and 4, a new method is proposed in Chapter 5 to estimate per-pixel classification confidence. A regression model is estimated to converts margin of victory into classification confidence by incorporating information from a separate test dataset. Results from Chapter 5 are evaluated in four aspects: (1) exploratory data analysis (EDA) using bi-histogram of PPCC for correct and error pixels; (2) statistical relationship between binned PPCC (BPC) with PPCC using full coverage reference data; (3) stability of PPCC estimates; (4) impact of sample design on PPCC estimates. There are several significant findings from Chapter 5 that need mention here. First, stable estimation of PPCC can be obtained using margin of victory. Second, sample size is the most important factor influencing PPCC. The estimate of PPCC is robust when sample size is above certain level (in my case, 1860 pixels). When sample size is too small, no reliable PPCC estimates are available. The impact of sampling schemes on PPCC estimates is negligible. The estimated PPCCs are comparable across classifiers and thus indicate the performance of different classifiers.

2 Implications

This dissertation focuses on classification confidence at pixel level. It falls into the context of spatial data quality which is an important topic both in geographical information science and remote sensing. With the development of GIS and remote
sensing, the importance of spatial data quality has attracted more and more attention from academic community, governmental agency, and industry. This study will contribute to the spatial data quality field in at least three points.

First, the comprehensive literature review revealed the current status in per-pixel classification confidence. Specifically, this dissertation classified previous methods into three categories. The advantages and limitations for the methods in each category are discussed.

Second, the evaluation of previous methods pointed out the directions on what can be done and what cannot. The evaluation found that interpolation based method performed poor due to the interpolation effect. This finding implied that interpolation methods have to comply with the spatial distribution pattern of the variables to be interpolated.

Third, the estimated per-pixel classification confidence can be applied to further studies such as error propagation. For example, by integrating the spatial distribution of classification confidence in land use land cover change detection, the map users will be well informed of the areas that are of high or low data quality.

3 Further studies

There are several interesting fields that may be examined in the future.

First, more studies can be carried out on method evaluation. In this dissertation, I did not evaluate regression based method due to the lack of data. This remains a topic for future studies.

Second, this study is carried out based on three remote sensing images with complete coverage reference maps. As a methodology study, it is best if the method can be tested
on as many cases as possible. Due to the limited availability of reference data, it is not practical to test the methods on as many real remote sensing images as possible. However, simulated data can solve this problem. In the future, the author plans to evaluate and test the methods using simulated data.

Third, this dissertation estimated classification confidence using classification scores only. While the results look good, it is possible to improve the method or develop new methods by combining different techniques together. For example, new methods may be developed by combining bootstrap method classification score based method.
References


Burnicki, A.C. (2012). Impact of error on landscape pattern analyses performed on land-cover change maps. Landscape Ecology, 27, 713-729


171


173


Platt, J. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schölkopf & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 61-74)


